
# Addis Ababa University
# College of Natural Science

## Department of Mathematics

**Dimensionality Reduction and Classification using Improved Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)**

**A Thesis Submitted to Addis Ababa University, College of Natural Science, in Partial Fulfillment of the Requirements for the Degree of Master's of Science in Optimization**

**By:**
**Endale Deribe Jiru**

**Advisor:**
**Berhanu Guta (PhD)**

*June, 2020*

*Addis Ababa*

# Declaration

I declare that this research paper entitled "Dimensionality Reduction and Classification using Improved Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)" is my original work and has not been used by others for any other requirements in any other university and all sources of information in the study has been appropriately acknowledged.

Endale Deribe Jiru_____

Date _____

## Approval Sheet

Dimensionality Reduction and Classification using Improved Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

By

Endale Deribe Jiru

**Approved by:**

Advisor: Berhanu Guta (PhD) _____

Date: _____

Examiner 1: _____ Signature _____

Examiner 2: _____ Signature _____

*Dedicated to*

*Deribe Jiru (my father) and Diso Aredo (my grandmother)*

## Acknowledgements

First of all, I would like to express my sincere thanks and gratitude to my respected advisor Dr. Berhanu Guta for his knowledgeable comment and suggestions to make this thesis work real. Secondly, I would like to thank and appreciate Dr. Mesfin Redi for his invaluable and moral support to start my study earlier and providing suggestions and support until the completion of this work.

Next, I would like to thank my uncle Mr. Dechasa Jiru (Dechure) for his encouragement and support throughout my study at the university. I would like to express my heartfelt thanks to Mr. Elias Redi and all Hajji Redi Abshiro families for their moral, continuous encouragement and support until the completion of my study.

I wish to express my sincere thanks to my mother W/ro Tewabech Kura, all my sisters and brothers and families for their motivation and lovely support in every step of my life. Last but not least, I would like to thank my respected wife W/ro Roman Redi and my lovely daughter Hawi Endale for all the joy, happiness and others they brought into my life.

# Abstract

*Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) are two popular methods for dimensionality reduction. PCA is a multivariate data analysis method, which uses an orthogonal transformation to convert a set of possibly correlated observations into a set of linearly uncorrelated components called principal components, whereas LDA, is a method to find a linear combination of observations which separates two or more classes of objects by finding a low dimensional subspace that keeps data points from different classes far apart and those from the same class as close as possible. In this study, dimensionality reduction and classification were performed using improved PCA and LDA in order to identify the most important discriminant variables (CGAs) from the phenolic compounds content dataset of the green coffee beans for the purpose of identifying their geographical origin. The dataset used in this work were extracted from published article (Mehari, B. et al., 2016) by applying the Box-Muller method on the mean and standard deviation values of the green coffee beans given for each regional and sub-regional category. Prior to constructing PCA model, the suitability of dataset was assessed using Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of Sphericity. Subsequently, the dataset were subjected to principal component analysis (PCA) with Varimax rotation method to identify the most discriminating compound corresponding to the green coffee beans and LDA model was developed to classify the coffee beans. The findings of this work showed 3-caffeoylquinic acid (3-CQA), 4,5-dicaffeoylquinic acid (4,5-diCQA), 3,5-dicaffeoylquinic acid (3,5-CQA) to 4,5-dicaffeoylquinic acid (4,5-diCQA) concentration ratio, and 4,5-dicaffeoylquinic acid (4,5-diCQA) to 3,4-dicaffeoylquinic acid (3,4-diCQA) concentration ratio were identified as the most discriminating compounds for the authentication of the various regional green coffee beans. Among these, 3-CQA and 4,5-diCQA were selected as suitable discriminant marker compounds for green coffee beans originating from Northwest (Benishangul and Finoteselam) and East (Harar) studied regions, respectively, both at regional and sub-regional levels. Moreover, at sub-regional level, sample of coffee beans from Jimma A, Wollega, and Sidama SA were distinguished by the 3,5-diCQA to 4,5-diCQA concentration ratio while the 4,5-diCQA to 3,4-diCQA concentration ratio was found appropriate to differentiate coffee beans from Yirgachefe and Jimma-B from the other coffee varieties. The results of LDA were in line with the PCA results, indicating that the LDA model was able to classify almost all of the coffee beans accurately based on the their geographical origin. The recognition and prediction abilities of the LDA model were 94% and 92.4%, respectively, at the regional level and 94.3% and 93.3%, respectively, at the sub-regional level and hence, best discrimination of green coffee beans was achieved both at regional and sub-regional. Further, comparisons between results obtained in this work and provided in the literature demonstrate the superiority of the improved methods.*



**Key words:** Eigenvalue, Eigenvector, Dimensionality reduction, Classification, Principal component analysis, Linear discriminant analysis, Coffee beans.

# Table of Contents

**Contents** **Page**

# List of Tables

**Page**

# List of Figures

**Page**

## Abbreviations

| | |
|---|---|
| **BTS** | Bartlett's Test of Sphericity |
| CGA | Chlorogenic Acid |
| DA | Discriminant Analysis |
| DR | Dimensionality Reduction |
| KMO | Kaiser-Meyer-Olkin |
| KDA | Kernel Discriminant Analysis |
| KPCA | Kernel Principal Component Analysis |
| LDA | Linear Discriminant Analysis |
| **MSA** | Measure of Sampling Adequacy |
| **PC** | Principal Component |
| PCA | Principal Component Analysis |
| RFE | Recursive Feature Elimination |
| SPSS | Statistical Package for Social Science |
| SVD | Singular Value Decomposition |

# CHAPTER

# 1 INTRODUCTION

## 1.1 Background (Motivations) of the Thesis

### 1.1.1 Dimensionality Reduction and Classification

Technological advancement and innovations have brought massive high dimensional data called Big Data. This has encouraged advancement of computational techniques considering the major issues of massive dataset related to the major challenges like volume, velocity, and variety, which are mainly related to dimensions (Kpigigbue N-Aabe et al, 2019). Due to "the curse of dimensionality", many data reduction techniques become weak when implemented using high-dimensional data. Even though data points of large number used, they remain scatter and be out of sight in the space of high-dimension, which is practically impossible to explore and analyze (Holmes S. and Huber W., 2019). So, by applying dimensionality reduction techniques on those data, these challenging and troublesome situations can be alleviated. Thus, the transformation of high dimensional data to low-dimensional data, which retain the signal of interest, while removing noise can be instrumental in to understand the patterns of data and hidden structures (Nguyen LH and Holmes S, 2019).

The underlying assumption for dimensionality-reduction techniques is that keeping the most useful information of the original high-dimensional datasets in a low-dimensional transformed subspace. Hence, the main goal of dimensionality reduction techniques is to get accurate and easily understandable representation of the original dataset with the removal of statistically redundant information (Bishop, 2006; Jolliffe and Cadima, 2016). Due to the needs of highly computing speed and availability of enough storage space, applying dimensionality reduction and analyzing the structure of high-dimensional data by visualization (mapping to a lower dimension, i.e. using two or three dimensions) is an important issue in different research areas.

Principal component analysis (PCA) and Linear Discriminant Analysis (LDA) are two popular methods for dimensionality reduction and data visualization (Bishop, 2006). PCA is a data analysis method, which uses an orthogonal transformation to convert a set of possibly correlated observations into a set of linearly uncorrelated components called principal components (Jolliffe, 2002). On the other hand, LDA, a generalization of Fisher's linear

discriminant, is a method to find a linear combination of observations which identifies or separates two or more categories of objects by finding a low dimensional subspace that keeps data points from different classes far apart and those from the same class as close as possible (Bishop, 2006). In other words, PCA and LDA are two widely used multivariate statistical methods used for dimensionality reduction and classification in the areas of pattern recognition, and science and engineering applications (Bishop, 2006; Charu C. Aggarwal, 2014; A. Tharwat, 2016).

Recently, several authors have developed and applied different dimensionality reduction and classification algorithms. Gupta et al. (2002) introduced dimensionality reduction techniques that improve the performance of classification algorithms. They compared the linear and non-linear methods for image recognition. They revealed that techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Kernel Principal Component Analysis (KPCA), and Kernel Discriminant Analysis (KDA) were used to reduce the dimensionality of original dataset and then classification was performed on the reduced dataset using support vector machine and nearest neighbor.

Lei and Govindaraju (2005) introduced PCA and Recursive Feature Elimination to speed up multiclass support vector machine. They revealed that PCA outperformed RFE. Similarly, Ivosev et al. (2008) discussed the dimensionality reduction and visualization techniques with the help of PCA and they showed that the clustering of variables using principal component were used in many applications, including tissue imaging, xenobiotic metabolism and metabolomics, which enabled them to reduce the dimensionality and improve visualization techniques on high dimensional data. Dash et al. (2010) constructed a hybridized Rough-PCA for data classification. By applying the Rough-PCA to the original dataset, they discovered the most discriminate variables, which could sufficiently classify the given dataset. Finally, they came with techniques that allowed them to obtain a minimal subset of features, which retain a high accuracy in the original attributes representation.

Takeda et al. (2012) proposed a unified robust classification model that optimizes the existing classification models such as fisher discriminant analysis named Linear Discriminant Analysis. They provided well-defined theoretical results, improved the existing techniques and investigated relationships among different classification models. Despite the development of PCA models that applied for dimensionality reduction, there are few studies focused on the retention outcome of key information of from the original dataset in classical

PCA (Hosseini, H.M. and Kaneko, S.; 2011). Study by Shang, L. and Wang, S. (2014) introduced improved classical PCA method. However, they couldn't implement it to different applications but limited its application to positive values. Therefore, this study tries to improve the classical PCA to expand the value of application.

Data normalization is a preprocessing technique for dimensionality reduction and classification to transform the original dataset into a desired range, which improves the data quality and reduces the discrepancies in the original dataset in order to make them fit for use. Data preprocessing is an important step in dimensionality reduction and data classification, which improves the performance of the techniques and algorithms. Even though many data analysis tools are developed and easily accessible, there is needs and requirement for data standardization techniques, which analyze the data intelligently. Thus, data normalization (standardization) removes the inconsistency and ambiguity in the datasets (Rathod and Momin, 2012). Chandrasekhar et al. (2011) studied the different data normalization and clustering techniques to eliminate missing values and avoid the redundant values in the original gene expression dataset.

In most dimensionality reduction techniques, the primary step is identifying whether the selected variables are interrelated to each other, since information overlap could make evaluation results biased (Shirali et al., 2016). Taking into account the fact that information overlap between variables is eliminated by applying PCA (Shirali et al., 2016; Jolliffe, I.T., 2002), the current study aimed to introduce principal component analysis as dimensionality reduction technique to examine the interrelations between variables of the green coffee beans dataset.

Studies show that PCA models have been commonly implemented using original dataset for dimensional reduction (Liu, B.S. et al., 2014; Coussement, A. et al., 2016; Rajesh, S. et al., 2018). These original datasets contains key information mainly in the two aspects (Shang, L.; Wang, S., 2014): the distribution information (or spread of data) among all variables, which is reflected by the variance; and the significant or insignificant relationships between the variables of the given dataset, which is reflected by correlation coefficient matrix (Shang, L.; Wang, S., 2014).

In order to apply normalization (standardization) on the original dataset having large differences in the measured scales, essential consideration should be taken for PCA in order

to avoid the loss of those key information Shang and Wang, 2014). However, the normalization (standardization) process in classical PCA makes the variances of all indicators equal (i.e. equal to 1), which eliminates the information of dispersion degree contained in the given dataset (Hao, R.X. et al., 2013; Shang and Wang, 2014).

This study attempts to use improved PCA and LDA modeling with a new normalization (standardization) method to perform dimensionality reduction and to identify most discriminating compounds useful for the differentiation of the various green coffee beans.

The significance of this study lies in following three aspects: (1) proposing improved PCA that allowed the standardized dataset retain more key information; (2) proposing LDA model that can result in better classification accuracy of the green coffee beans; and (3) proposing a multi-dimensional perspective that can contribute better understanding of the authentication of the green coffee beans based on their geographical origin.

### 1.1.2 Green Coffee Beans

Coffee is the second most important commodity in international trade according to the volume of trade. Coffee trade is international that involves networked trade covering developing and developed countries, which are its main consumers. The price of coffees in the international market, which depends on the quality of the coffee beans, has a direct correlation with the taste of the final consumed product. In addition to coffee quality, its originality and traceability seen as important factors in global trade, and hence, the identification and determination of quality and originality of coffee is necessary. (F Kurniawan et al., 2019).

Coffee, which cultivated in different regions of Ethiopia, plays a vital role in the country's economy and become a major source of foreign exchange earnings. There is a price difference for Ethiopian coffee products, which depends on the region of production (or origin), is determined by their flavors. This can be seen as the cause for the adulteration of expensive varieties with cheaper coffee varieties and fraud regarding to the production (or origin) areas within the country (Mehari, B. et al., 2016).

Authentication of coffee origin (production country or region), as well as coffee varieties, is highly demanded by international consumers as additional attribute of quality, and thus

consumers being willing to pay attractive prices for coffee varieties from particular areas (or regions of origins). In this context, effective and reliable methods to prevent fraudulent practices become necessary. For that purpose, the aim of this work focused on the Principal Component Analysis (PCA) and Linear Discriminant Analysis (DA) that allow identification of the most discriminating factors to distinguish the green coffee beans and address the classification and authentication of coffee samples based on their geographical origin.

### 1.1.3 Principal Component Analysis (PCA)

The reduction of high dimensional data comprised of large number of interrelated variables, while retaining as much as possible of the variation present in the data set is the core brain behind the application of principal component analysis (PCA) (I.T. Jolliffe, 2002). The idea behind PCA is the challenge of identifying patterns in dataset. Principal component takes charge of highlighting similarities and differences in a high dimensional data to meet graphical representation, and is hinged behind eigenvalues and eigenvectors making PCA a powerful technique used in data analysis (Lindsay I Smith, 2002).

Using mathematical projection, the original dataset, which may have involved many variables, can often be interpreted in just a few variables (i.e. *the principal components*). The central idea of principal component analysis is to reduce the dimensionality of a dataset in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This reduction is achieved by transforming to a new set of variables, the PCs, which are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables (Mishra, S. et al., 2017).

Principal Component analysis (PCA) is a technique, which is used in data analysis in different areas ranging from image recognition to many scientific applications. The increasing amount of data to be analyzed has increased the computational costs in drawing useful conclusions from it. PCA is, thus, used to lower the dimensionality of this big data to reduce the cost and improve data visualization (Jian Vora, 2019).

Mathematically, given n vectors in $R^{\mathbf{d}}$, PCA deals with finding the k-dimensional subspace which captures the maximum variance of the data. According to Jolliffe (2002), it extracts the principal components that preserve most of the variation in the original dataset. Let $X$ be a

random vector in $\mathbb{R}^d$ with mean zero and covariance matrix **S**. Then, PCA needs projection direction vectors, $v_1, v_2, \ldots, v_k \in \mathbb{R}^d$, such that

$$v_1 \in \text{argmax}_{\|v\|_2=1} v^T \mathbf{S} v,\ v_2 \in \text{argmax}_{\|v\|_2=1, v \perp v_1} v^T \mathbf{S} v,$$
$$v_3 \in \text{argmax}_{\|v\|_2=1, v \perp v_1, v_2} v^T \mathbf{S} v, \ldots$$

In other words, $\{v_j\}_{j=1}^k$ are the top *k* eigenvectors of **S**. Given $V_k \equiv (v_1, \ldots, v_k)$, hence, to achieve the goal of dimensionality reduction, we can then project the given high dimensional data onto the low dimensional space spanned by columns of $V_k$. Since $V_k$ captures the most variation in the dataset, these projected data points approximately preserve the geometric properties of the original data.

### 1.1.4 Linear Discriminant Analysis (LDA)

Linear discriminant analysis produces an uncorrelated number of dataset, which can be maximized for class separation. The discriminant functions effectively project one column of dataset onto the other and for this to be done, the target column has to be selected (Kpigigbue N-Aabe et al, 2019). LDA chooses features that maximize the ratio of classes and spread into the class. The LDA procedure aims to find an optimal projection so that it can project the given data in a space with a smaller dimension where all patterns can be classified as much as possible (Hapsari and Syamsuryadi, 2019).

In LDA algorithm, we first analyze between-class and within-class scatter measures used in discriminant analysis. Accordingly, LDA seeks to maximize the average pair-wise distance between class means and minimize the average within-class pair-wise distance over all classes. Moreover, LDA computes an optimization problem, which simultaneously maximizes the between-class scatter measure and minimizes the within-class scatter measure both at once.

So, by calculating between-class variance ($S_B$) and within-class variance ($S_W$), the LDA's transformation matrix (W) can be calculated using the eigen-vector equation:

$$S_W W = \lambda S_B W$$

where λ represents the eigenvalues of the transformation matrix (W). Thus, the solution of above problem can be obtained by finding the eigenvalues ($\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$) and

eigenvectors ($V = \{V_1, V_2, \dots, V_M\}$) of $W = S_W^{-1} S_B$, if $S_W$ is non-singular ( Lu et al., 2003; Ye et al., 2004).

The eigenvalues are scalar values, while the eigenvectors are non-zero vectors, which provide us with the information about the LDA space. The eigenvectors represent the directions of the new LDA space, and the corresponding eigenvalues represent the scaling factor, or the magnitude of the eigenvectors (Zhu and Ogihara, 2006). Thus, each eigenvector represents one axis of the LDA space, and the associated eigenvalue represents the strength of this eigenvector.

## 1.2 Problem Statement

Recent advancement in technologies resulted in an exponential growth in dataset with respect to sample size (observations) as well as dimensions (variables). The efficient and effective administration of such huge data creates challenges for the users. Retrieving information manually from such huge amount of datasets becomes impractical (Adnan Ullah, et al., 2017). These datasets may also have noisy, irrelevant or redundant features, which pose a challenge for the researchers to automatically extract useful information, knowledge and structure (pattern).

Original high-dimensional data often contain measurements on uninformative or redundant variables. Dimensionality reduction is frequently used for data compression, exploration, and visualization. Although many dimensionality reduction techniques have been developed and implemented in standard data analytic pipelines, they are easy to misuse, and their results are often misinterpreted in practice (Nguyen LH and Holmes S, 2019). Existing dimensionality reduction and classification frameworks lack in producing easily interpretable features, understandable patterns and interestingness results, (Geng and Hamilton, 2006).

Principal Component Analysis (PCA) is a novel way of dimensionality reduction. This problem essentially boils down to finding the top k eigenvectors of the data covariance matrix. PCA is essentially finding a k-dimensional subspace which captures the maximum variance of the data. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be n vectors in $\mathbb{R}^d$. We assume that they are mean-centered. The goal of PCA is to reduce the dimensionality of these vectors using a linear transformation. The matrix $V \in \mathbb{R}^{k \times d}$, where $k < d$, induces a mapping $\boldsymbol{x} \rightarrow \boldsymbol{V}\boldsymbol{x}$, where $V\boldsymbol{x}$ is the lower dimensional representation of $\boldsymbol{x}$. A second matrix $U \in \mathbb{R}^{d \times k}$can be used to

approximately recover each original vector $\boldsymbol{x}$ from its compressed (reduced) form. That is, for a compressed vector $\boldsymbol{y} = \boldsymbol{Vx}$, where $\mathbf{y}$ is in the low-dimensional space $\mathbb{R}^k$, we can construct $\tilde{\boldsymbol{x}} = U\boldsymbol{y} = UV\boldsymbol{x}$, so that $\tilde{\boldsymbol{x}}$ is the recovered version of $\boldsymbol{x}$ in the original space $\mathbb{R}^d$. Thus, in PCA, the corresponding objective of finding the compression (reduction) matrix $V$ and the decompression (recovering) matrix $U$ is then phrased as optimization problem.

$f(U,V) = \sum_{i=1}^{n} \|x_i - UVx_i\|_2^2$ be objective function and the minimization problem is:

$$minimize_{V \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}} \sum_{i=1}^{n} \|x_i - UVx_i\|_2^2 \qquad (1.1)$$

That is, PCA tries to minimize the total squared distance between the original vectors and the recovered vectors.

Similarly, the goal of the LDA technique is to project the original data matrix onto a lower dimensional space that allows a good classification among different classes (groups). To achieve this goal, the first step is to calculate the separability between different classes (i.e. the distance between the means of different classes), which is called between-class variance or between-class matrix. Secondly, it needs to calculate the distance between the mean and the samples of each class, which is called within-class variance or within-class matrix. Following these two steps, the last step is to construct the lower dimensional space, which maximizes between-class variance and minimizes within-class variance (for justification see section 4.4). After calculating the between-class variance ($S_B$) and within-class variance ($S_W$), the transformation matrix (W) of the LDA technique can be calculated as in:

$$\text{maximize}_{\,W} \frac{W^T S_B W}{W^T S_W W} \qquad (1.2)$$

This formula can be reformulated as eigen-vector equation:

$$S_W W = \lambda S_B W$$

where $\lambda$ represents the eigenvalues of the transformation matrix (W). The solution of this problem can be obtained by calculating the eigenvalues ($\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_M\}$) and eigenvectors ($V = \{V_1, V_2, \dots, V_M\}$) of $W = S_W^{-1} S_B$.

Therefore, the aim of this work is to find solutions to the above optimization equations (1.1) and (1.2), which in turn make dimensionality reduction and classification on the given data matrix of dimensions (variables) and cases (observations).

## 1.3 Objectives of the Research Work

The main objective of this research work is to apply dimensionality reduction and Classification using improved PCA and LDA with the following specific objectives

- To develop an improved PCA model that can make the standardized dataset retain more key information.
- To develop an efficient LDA classification model with better recognition and prediction abilities.
- To allow better understanding of the authentication of the green coffee beans based on their geographical origin.

## 1.4 Organization of the Thesis

The structure of this thesis is organized as follows.

Chapter 1: Introduction- describes the background of the study, problem statement and objectives of the study. Chapter 2: Mathematical and Statistical Background- summarizes a few well-known results of linear algebra and statistics that will be exploited repeatedly in the thesis, and the optimization of dimensionality reduction presented. Then, Chapter 3: Literature Review- provides a brief overview of related works in relation to green coffee beans, normalization, dimensionality reduction and classification of high dimensional dataset. Chapter 4: Research Methods- discusses methods for dimension reduction and classification using PCA and LDA. Chapter 5: Results and Discussions, which presents experimental results and discussions of dimensionality reduction and classification using PCA and LDA. Finally, Chapter 6: Conclusions and Recommendations, provides conclusions and possible recommendations based on the findings of the study.

# CHAPTER

# 2 MATHEMATICS AND STATISTICS FOR DIMENSIONALITY REDUCTION

In order to make data analysis by Principal Component Analysis, we have to be thorough in matrix algebra and statistics. So, we discuss on Matrix Algebra by focusing on eigenvectors and eigenvalues, which are the fundamental principle to determine PCA and LDA, and also on Statistics which looks at distribution measurements, how the data is spread out. Moreover, the basic optimization techniques for dimensionality reduction are presented in this chapter.

## 2.1 Linear Algebra Background

We discuss the powerful method of principal component analysis (PCA) and linear discriminant analysis (LDA) in dimensionality reduction. Mathematically, PCA depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition (SVD) of rectangular matrices. It is determined by eigenvectors and eigenvalues. Eigenvectors and eigenvalues are numbers and vectors associated to square matrices. Together they provide the eigen-decomposition of a matrix, which analyzes the structure of this matrix such as correlation and covariance matrices (Mishra, S. et al., 2017). Computation of the PCs reduces to the solution of an eigenvalue-eigenvector problem for a positive-semi-definite symmetric matrix.

Linear algebra is the backbone of PCA development. Before go to PCA, understanding the algebraic implications is crucial. Let A be an m×n matrix of real numbers and $A^T$ its transpose. The following theorem is one of the most important theorems in linear algebra, which is the core point in Principal Component Analysis (PCA).

### 2.1.1 Diagonalization of Symmetric Matrices

**Definition 2.1.1:** *For a matrix A, if there exists a non-zero vector* $\boldsymbol{u}$ *and a real number* $\lambda$ *such that* $A\boldsymbol{u} = \lambda\boldsymbol{u}$*, then* $\lambda$ *is called an eigenvalue and* $\boldsymbol{u}$ *is called the corresponding eigenvector. If* $\lambda$ *is an eigenvalue, it is computed by solving the characteristic equation* $\det(A - \lambda I)=0$.

**Definition 2.1.2:** $A \in \mathbb{R}^{n\times n}$ *is a symmetric matrix if and only if* $A = A^T$.

**Theorem 2.1.1:** If A is symmetric, then any two eigenvectors from different eigenvalues are orthogonal.

***Proof:***

*Let* $\boldsymbol{v}_1$ *and* $\boldsymbol{v}_2$ *be two eigenvectors from two different eigenvalues* $\boldsymbol{\lambda}_1$ *and* $\boldsymbol{\lambda}_2$ *of* $A$.

*Let's show that* $\boldsymbol{v}_1.\boldsymbol{v}_2 = 0$.

$$\lambda_1(\boldsymbol{v}_1.\boldsymbol{v}_2) = (\lambda_1\boldsymbol{v}_1)^T\boldsymbol{v}_2 = (\boldsymbol{A}\boldsymbol{v}_1)^T\boldsymbol{v}_2 \quad \text{(By Definition of eigenvector)}$$

$$= v_1^T\boldsymbol{A}^T\boldsymbol{v}_2 \quad \text{(Transpose of product)}$$

$$= v_1^T(\boldsymbol{A}\boldsymbol{v}_2) \quad \text{(A is symmetric, } i.e.\ \boldsymbol{A}^T = \boldsymbol{A}\text{)}$$

$$= v_1^T(\lambda_2\boldsymbol{v}_2) \quad \text{(By Definition of eigenvector)}$$

$$= \lambda_2(\boldsymbol{v}_1.\boldsymbol{v}_2)$$

Thus, $\lambda_1(\boldsymbol{v}_1.\boldsymbol{v}_2) = \lambda_2(\boldsymbol{v}_1.\boldsymbol{v}_2) \Longrightarrow (\lambda_1 - \lambda_2)(\boldsymbol{v}_1.\boldsymbol{v}_2) = 0$.

Since $\lambda_1 - \lambda_2 \neq 0$ *(i.e.* $\lambda_1 \neq \lambda_2$), consequently, $\boldsymbol{v}_1.\boldsymbol{v}_2 = 0$.

**Theorem 2.1.2:** If matrix A is symmetric, then all its eigenvalues are real.

***Proof:***

Suppose $v, w$ two vectors of matrix *A*. Note that the dot product can be extended to complex vectors as $(v, w) = \sum_i \overline{v}_i w_i$. For real vectors it satisfies $(v, w) = v.w$ and has the property $(Av, w) = (v, A^T w)$ for real matrices *A* and $(\lambda v, w) = \overline{\lambda}(v, w)$ as well as $(v, \lambda w) = \lambda(v, w)$.

Now $\overline{\lambda}(v, v) = (\lambda v, v) = (Av, v) = (v, A^T v) = (v, Av) = (v, \lambda v) = \lambda(v, v)$. This shows that $\overline{\lambda} = \lambda$ because $(v, v) \neq 0$ for $v \neq 0$. Therefore, eigenvalues of a symmetric matrix are real.

**Definition 2.1.3:** *A matrix A is* ***diagonalizable*** *if there is an invertible matrix V such that A = VDV*$^{-1}$ *where D is diagonal matrix of eigenvalues of A.*

**Definition 2.1.4**: *A matrix A is* ***orthogonally diagonalizable*** *if and only if A = VDV*$^T$ *being V an orthogonal (V*$^{-1}$ *= V*$^T$*) and D is diagonal matrix of eigenvalues of A.*

**Theorem 2.1.3:** An n × n matrix A is orthogonally diagonalizable if and only if it is symmetric.

***Proof:***

Suppose A is orthogonally diagonalizable. This means

$A = VDV^{\mathrm{T}}$, where $V$ is orthogonal (so, $V^{\mathrm{T}} = V^{-1}$) and $D$ is diagonal matrix.

Transpose both sides gives $A^{\mathrm{T}} = (VDV^{\mathrm{T}})^{\mathrm{T}} = VD^{\mathrm{T}}V^{\mathrm{T}}$.

For a diagonal matrix $D$, we have $D = D^{\mathrm{T}}$, so we get $A^{\mathrm{T}} = VD^{\mathrm{T}}V^{\mathrm{T}} = VDV^{T} = A$, which implies $A^{\mathrm{T}} = A$.

Therefore, A is symmetric matrix.

**Theorem 2.1.4:** If a matrix $A \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ is symmetric, then $v^{T}Aw = w^{T}Av$ for all $v, w \in \mathbb{R}^{\mathrm{n}}$.

***Proof:*** Notice that since $v^{T}Aw$ is a scalar, it equals its own transpose. Thus,

$$v^{T}Aw = (v^{T}Aw)^{T} = w^{T}A^{T}v.$$

Since A is symmetric, we have $A^{T} = A$, and hence,

$$v^{T}Aw = w^{T}Av.$$

**Theorem 2.1.5:** Given a symmetric matrix A, the matrices $AA^{\mathrm{T}}$ and $A^{\mathrm{T}}A$ share the same nonzero eigenvalues.

***Proof:*** Let $\vec{v}$ be a (nonzero) eigenvector of $A^{T}A$ corresponding to eigenvalue $\lambda \neq 0$. Then,

$$(A^{T}A)\vec{v} = \lambda\vec{v}. \qquad (2.1.1)$$

Now, multiply both sides on the left by $A$, and group the parentheses as follows:

$$AA^{T}(A\vec{v}) = \lambda(A\vec{v}).$$

This implies that the vector $A\vec{v}$ is an eigenvector of $AA^{T}$ with eigenvalue $\lambda$. Moreover, from the first equation, if $A\vec{v}$ were zero, then $\lambda\vec{v}$ would be zero as well. However, we specifically said that $\vec{v} \neq \vec{0}$ and $\lambda \neq 0$, so this can't happen, and hence, $A\vec{v} \neq 0$. Therefore, the nonzero eigenvalue $\lambda$ of $A^{\mathrm{T}}A$ is also an eigenvalue of $AA^{T}$.

Similarly, if $\mu \neq 0$ is an eigenvalue of $AA^{T}$ corresponding to non-zero eigenvector $\vec{\omega}$,

$$(AA^T)\vec{\omega} = \mu\vec{\omega}. \qquad (2.1.2)$$

Now, multiply both sides on the left by $A^T$, and grouping:

$$A^T A(A^T\vec{\omega}) = \mu(A^T\vec{\omega}).$$

This implies that the vector $A^T\vec{\omega}$ is an eigenvector of $A^T A$ with eigenvalue $\mu$. Moreover, from (2.1.2), we have $(AA^T)\vec{\omega} = A(A^T\vec{\omega}) = \mu\vec{\omega}$. Since $\mu\vec{\omega} \neq 0$, so $A^T\vec{\omega} \neq 0$. Therefore, the nonzero eigenvalue $\mu$ of $AA^{\mathrm{T}}$ is also an eigenvalue of $A^T A$.

Hence, the matrices $AA^{\mathrm{T}}$ and $A^{\mathrm{T}}A$ share the same **nonzero** eigenvalues. □

**Theorem 2.1.6:** Given a symmetric matrix A, the eigenvalues of $AA^{\mathrm{T}}$ and $A^{\mathrm{T}}A$ are nonnegative numbers.

***Proof:*** Given a vector $\vec{v}$, we have $\|\vec{v}\|^2 = \vec{v}^T\vec{v}$.

Let $\vec{v}$ be an eigenvector of $A^T A$ corresponding to the eigenvalue $\lambda$. Then,

$$\|A\vec{v}\|^2 = (A\vec{v})^T(A\vec{v}) = \vec{v}^T(A^T A)\vec{v} = \vec{v}^T(\lambda\vec{v}) = \lambda\vec{v}^T\vec{v} = \lambda\|\vec{v}\|^2.$$

Since lengths are nonnegative, we see that $\lambda$ is nonnegative.

Similarly, replacing $A$ with $A^T$, we get the corresponding statement for $AA^T$.

### 2.1.2 Trace of Square Matrices

**Definition 2.1.5:** *The trace of an $n \times n$ matrix A is the sum of its diagonal entries. That is, if $A = [a_{ij}]$, then $trace(A) = \sum_{i=1}^{n} a_{ii}$.*

**Theorem 2.1.7:** For any two n × n matrices $A$ and $B$, we have $trace(AB) = trace(BA)$.

***Proof:***

Let $A = [a_{ij}]$, $B = [b_{ij}]$ be two $n \times n$ matrices such that $trace(A) = \sum_{i=1}^{n} a_{ii}$ and $trace(B) = \sum_{i=1}^{n} b_{ii}$. Then,

$trace(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki}$B

$$= \sum_{k=1}^{n}\sum_{i=1}^{n} b_{ki}a_{ik} \sum_{i=1}^{n}(BA)_{ii} = trace(BA)$$

**Theorem 2.1.8:** For an n × n symmetric matrix A, trace (A) is the sum of eigenvalues of A, that is, $\text{trace(A)} = \sum_{i=1}^{n} \lambda_i$.

***Proof:*** Since an $n \times n$ matrix A is symmetric, by theorem 2.1.3, it is orthogonally diagonalizable. That is, $A = VDV^T$, where $V$ is orthogonal matrix and $D$ is diagonal matrix. Then,

$$trace(A) = trace(VDV^T)$$

$= \sum_{i=1}^{n} v_i D v_i^T$ (by definition)

$= \sum_{i=1}^{n} [v_{ik}\lambda_k] v_i^T$, where $[v_{ik}\lambda_k]$ is a row vector with index k.

$$= \sum_{i=1}^{n}\sum_{k=1}^{n} v_{ik}\lambda_k v_{ik} = \sum_{k=1}^{n}\lambda_k \sum_{i=1}^{n} v_{ik}v_{ik}$$

$= \sum_{k=1}^{n} \lambda_k v_k * v_k$, where $v_k$ is a unit column vector.

$$= \sum_{k=1}^{n} \lambda_k$$

Thus, it shows that trace (A) is the sum of eigenvalues of A.

## 2.2 Statistics Background

In this section, the basic preliminary statistical tools for PCA and LDA were discussed.

### *Standard Deviation and Variance*

Standard Deviation of a set of observations of a series is the positive square root of the arithmetic mean of the squares of all the deviations from the arithmetic mean. Thus, in the calculation of standard deviation, first the arithmetic mean is calculated and the deviation of various items from the arithmetic mean is squared. Then the squared deviations are totaled and sum is divided by the number of items. Hence, standard deviation is a measure of dispersion of more mathematical significance. It is generally denoted by σ. The square of standard deviation is known as variance. So, variance is denoted by $\sigma^2$.

The formula for the standard deviation and variance for ungrouped data are:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})}{(n-1)}} \text{ and } Var(X) = \sigma^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})}{(n-1)}, \text{ respectively.}$$

Where, $\sigma$ = Standard Deviation,

$\sigma^2$ = Variance

$\bar{X}$ = Arithmetic mean,

$d = X - \bar{X}$ = Deviation of individual observation from arithmetic mean,

n = Number of observations, and,

$\sum d^2$ = Summation of squares of deviations.

### *Covariance*

Standard deviation and variance only operate on one dimension, so that we could only calculate the standard deviation for each dimension of the dataset independently of the other dimensions. However, it is useful to have a similar measure to find out how much the dimensions vary from the mean with respect to each other. Covariance is such a measure. Covariance is always measured between two dimensions. If we calculate the covariance between one dimension and itself, we will get the variance. The formula for covariance is very similar to the formula for variance. The formula for covariance with respect to variance is as follows.

$$Var(X) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})}{(n-1)}, \text{ and so, } Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)}$$

Where,

$\bar{X}$ = Arithmetic mean of data X

$\bar{Y}$ = Arithmetic mean of data Y

n = Number of observation

### *The Covariance Matrix*

Covariance is always measured between two dimensions. If we have a dataset with more than two dimensions, there is more than one covariance measurement that can be calculated. For example, from a 3 dimensional dataset (dimensions X, Y, Z) we could calculate the $Cov(X,Y)$, $Cov(Y,Z)$ and $Cov(X,Z$. In fact, for an n-dimensional dataset, we can calculate $\frac{n!}{(n-2)!*2}$ different covariance values.

A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix. So, by definition the covariance matrix for a set of data with n-dimensions is as follows.

$$C^{M*N} = (c_{ij},\ c_{ij} = Cov(Dim_i, Dim_j)$$

Where,

$C^{M*N}$ is a matrix with m rows and n columns, and

$Dim_k$ is the k$^{th}$ dimension.

This typical formula says that if you have an n-dimensional dataset, then the matrix has n rows and columns (so is square) and each entry in the matrix is the result of calculating the covariance between two separate dimensions. For example, the covariance matrix for an imaginary 3 dimensional dataset, using the usual dimensions X, Y and Z, has 3 rows and 3 columns, and the values are:

$$\begin{bmatrix} Cov(X,X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Cov(Y,Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Cov(Z,Z) \end{bmatrix}$$

Notice that the diagonal elements are the covariance value between one of the dimensions with itself. These are the variances for that particular dimension. In addition, the covariance matrix is symmetrical about the main diagonal, $Cov(X,Y) = Cov(Y,X)$.

## 2.3 Optimization for Dimensionality Reduction

**Definition 2.3.1:** *Given a data matrix of dimension d and observations (cases) n, Dimensionality Reduction is the task to find a k-dimensional representation of a d-dimensional dataset, with k < d such that the d-dimensional information is maximally preserved.*

The main motivations to perform dimensionality reduction are uncovering the intrinsic dimensionality of the data, data visualization, reduction of redundancy and noise, and computational or memory savings (Mohammed J. Zaki and Wagner Meira Jr., 2014).

### 2.3.1 Principal Component Analysis

The main tasks of this section are to make understanding of what a PC is and why PCA can be interpreted as a data reduction technique, describe the link between PCA and the eigenvectors and eigenvalues of the covariance matrix, how to compute PCA efficiently for dataset of dimension (d) and number of observations (n) with d<n. In addition, ways to determine the number of PCs and how PCA captures the major axis of variation in the dataset are described in details.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n vectors in $\mathbb{R}^d$. We assume that they are mean-centered. The goal of PCA is to reduce the dimensionality of these vectors using a linear transformation. The matrix $V \in \mathbb{R}^{k \times d}$, where $k < d$, induces a mapping $\boldsymbol{x} \to \boldsymbol{V}\boldsymbol{x}$, where $V\boldsymbol{x}$ is the lower dimensional representation of $\boldsymbol{x}$. A second matrix $U \in \mathbb{R}^{d \times k}$can be used to approximately recover each original vector $\boldsymbol{x}$ from its compressed (reduced) form. That is, for a compressed vector $\boldsymbol{y} = \boldsymbol{V}\boldsymbol{x}$, where **y** is in the low-dimensional space $\mathbb{R}^k$, we can construct $\tilde{\boldsymbol{x}} = U\boldsymbol{y}$, so that $\tilde{\boldsymbol{x}}$ is the recovered version of $\boldsymbol{x}$ in the original space $\mathbb{R}^d$.

In PCA, the corresponding objective of finding the compression (reduction) matrix *V* and the decompression (recovering) matrix *U* is then phrased as optimization problem.

$f(U,V) = \sum_{i=1}^{n} \|x_i - UVx_i\|_2^2$ be objective function and the minimization problem is:

$$minimize_{V \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}} \sum_{i=1}^{n} \|x_i - UVx_i\|_2^2 \qquad (2.3.1)$$

That is, PCA tries to minimize the total squared distance between the original vectors and the recovered vectors.

**Theorem 2.3.1:** Let (U,V) be a solution to the objective function (2.3.1). Then the columns of U are orthonormal (that is, $U^TU=I$) and $U = V^T$.

***Proof:***

We make the following assumptions:

Choose any *U* and *V* and consider the mapping $\boldsymbol{x} \to \boldsymbol{U}V\boldsymbol{x}$.

The range of this mapping, $R = \{UV\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^d\}$ is an *k-dimensional* linear subspace of $\mathbb{R}^d$.

Let $W \in \mathbb{R}^{d\times k}$ be a matrix whose columns (i.e. $w_1, \dots, w_k$) form an orthonormal basis of this subspace (i.e. $W^TW = I$).

Hence, for each $x_i$, there is $y_i \in \mathbb{R}^k$ such that $UVx_i = Wy_i$.

Each vector in $\mathbb{R}$ can be written as $W\boldsymbol{y}$ where $\boldsymbol{y} \in \mathbb{R}^k$.

For every $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^k$, we have

$$f(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} - W\boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|^2 - \boldsymbol{x}(Wy)^{\boldsymbol{T}} - (Wy)^{\boldsymbol{T}}\boldsymbol{x} + (Wy)^{\boldsymbol{T}}Wy.$$

$$= \|\boldsymbol{x}\|^2 - \mathbf{2}y^{\boldsymbol{T}}W^{\boldsymbol{T}}\boldsymbol{x} + \boldsymbol{y}^{\boldsymbol{T}}W^{\boldsymbol{T}}Wy$$

$$= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \mathbf{2}\boldsymbol{y}^{\boldsymbol{T}}(W^{\boldsymbol{T}}\boldsymbol{x}) \qquad (2.3.2)$$

Then minimizing $f(x,y)$ with respect to y gives:

$\nabla_y f(x,y) = 2y - 2W^{\boldsymbol{T}}x = \mathbf{0}$, which implies $y = W^Tx$.

So, (2.3.2) is minimized for $y = W^Tx$.

Therefore, for each $\boldsymbol{x_i}$ we have that:

$f(U,V) = \sum_{i=1}^n \|x_i - UVx_i\|_2^2 = \sum_{i=1}^n \|x_i - Wy_i\|_2^2 \geq \sum_{i=1}^n \|x_i - WW^Tx_i\|_2^2$.

As U, V are optimal, so we have

$$\sum_{i=1}^n \|x_i - UVx_i\|_2^2 = \sum_{i=1}^n \|x_i - WW^Tx_i\|_2^2 \qquad (2.3.3)$$

Thus, we can replace $U$ *by* $W$ and $V$ *by* $W^T$

Therefore, $U = V^T$ and $U^TU = I$. ■

From the above theorem, we have $V = U^T$and $U^TU = I$. So, we can rewrite the optimization problem (2.3.1) as:

$$\text{minimize}_{U\in\mathbb{R}^{d\times k}:\, U^TU=I} \sum_{i=1}^n \|x_i - UU^Tx_i\|_2^2 \qquad (2.3.4)$$

**Claim:** $\|x - UU^Tx\|^2 = \|x\|^2 - trace(U^Txx^TU)$.

***Proof:***

For every $x \in \mathbb{R}^d, U \in \mathbb{R}^{d\times k}$ with $U^TU = I$,

$$\begin{aligned}\|x - UU^T x\|^2 &= trace[(x - UU^T x)^T (x - UU^T x)] \\ &= trace[(x^T - x^T UU^T)(x - UU^T x)] \\ &= trace[x^T x - x^T UU^T x - x^T UU^T x + (x^T UU^T)(UU^T x)] \\ &= trace[x^T x - 2x^T UU^T x + (x^T UU^T x)], \text{ (i.e. } U^T U = I) \\ &= trace[x^T x - x^T UU^T x] \\ &= trace[x^T x] - trace[x^T UU^T x] \\ &= \|x\|^2 - trace(U^T x x^T U) \qquad (2.3.5)\end{aligned}$$

Thus, using (2.3.5), we can reformulate the minimization problem (2.3.4) as a trace maximization problem:

$$\text{maximize}_{U \in \mathbb{R}^{d \times k} : U^T U = I} \, trace(U^T \textstyle\sum_{i=1}^{n} x_i \, x_i^T U) \qquad (2.3.6)$$

Let $\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$, then the above optimization problem becomes:

$$\text{maximize}_{U \in \mathbb{R}^{d \times k} : U^T U = I} \, trace(U^T \mathbf{S} U) \qquad (2.3.7)$$

**Theorem 2.3.2:** Let $\mathbf{S} = VDV^T$ be the spectral decomposition of **S**. $D$ is a diagonal matrix, such that $D_{i,i}$ is the i[th] largest eigenvalue of **S.** The columns of $V$ are the corresponding eigenvectors, and $V^T V = VV^T = I$. Then the solution of (2.37) is the matrix $U$ whose columns are the $k$ first eigenvectors of **S**.

***Proof:***

Choose a matrix $U \in \mathbb{R}^{d \times k}$ with orthonormal columns and let $B = V^T U$.

Then, $VB = VV^T U = U$. From this, we have

$$U^T \mathbf{S} U = B^T V^T VDV^T VB = B^T DB \qquad (2.3.8)$$

and hence,

$$trace\,(U^T \mathbf{S} U) = \textstyle\sum_{j=1}^{d} D_{jj} \sum_{i=1}^{k} B_{ji}^2 \qquad (2.3.9)$$

Since $B = V^T U$, then $B^T B = U^T VV^T U = U^T U = I$.

Hence, the columns of B are orthonormal and $\sum_{j=1}^{d} \sum_{i=1}^{k} B_{ji}^2 = k$.

Now, define the matrix $\tilde{B} \in \mathbb{R}^{d \times d}$ such that the first $k$ columns are the columns of $B$ and moreover, $\tilde{B}^T \tilde{B} = I$. Then for every $j$, $\sum_{i=1}^{d} \tilde{B}_{ji}^2 = 1$, which implies that $\sum_{i=1}^{k} \tilde{B}_{ji}^2 \leq 1$.

It follows that

$$trace\ (U^T \mathbf{S} U) \leq \underset{\beta \in [0,1]^d : \|\beta\|_1 \leq k}{maximize} \sum_{j=1}^{d} D_{jj}\, \beta_j = \sum_{j=1}^{k} D_{jj}$$

Hence, for every matrix $U \in \mathbb{R}^{d \times k}$ with orthonormal columns, (i.e. $U^T U = I$),

$trace\ (U^T \mathbf{S} U) \leq \sum_{j=1}^{k} D_{jj}$.

So, if we set *U* to the matrix with the first *k* leading eigenvectors of **S** as its columns, we get $\mathbf{trace}\ (\mathbf{U^T S} U) = \sum_{j=1}^{k} D_{jj}$ and hence, the optimal solution. ■

**How to choose the number of principal components?**

The principal components should capture *p* per cent (i.e. *p* is usually >= 80%) of the total variance in the data. Total variance in the data is $\sum_{i=1}^{d} \lambda_i$, and total variance captured by first *k* eigenvectors is $\sum_{i=1}^{k} \lambda_i$. Hence, the variance explained (*p)* is the ratio between the two, that is $p = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$.

**Theorem 2.3.3:** The variance captured by the first k eigenvectors of S is the sum over its k largest eigenvalues, *that is* $\sum_{i=1}^{k} \lambda_i$.

***Proof:*** The variance in a dataset *X* is defined as:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i - 0\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \langle x_i, x_i \rangle$$

$$= trace\ (\mathbf{S}) = trace\ (V^T D V) = trace\ (V V^T D) = trace\ (D)$$

$$= \sum_{i=1}^{d} D_{ii} = \sum_{i=1}^{d} \lambda_i. \qquad (17)$$

The variance in a projected dataset *Y= VX*, with $V = [v_1, \ldots, v_k]^T$, is defined as

$$Var(Y) = Var(VX) = \frac{1}{n}\sum_{j=1}^{n} \left\| V x_j - 0 \right\|^2 = \frac{1}{n}\sum_{j=1}^{n} \left\| V x_j \right\|^2$$

$$= \frac{1}{n}\sum_{j=1}^{n} \langle V x_j, V x_j \rangle = \frac{1}{n}\sum_{j=1}^{n} x_j^T V^T V x_j$$

$$= \frac{1}{n}\sum_{j=1}^{n} x_j^T (v_1 v_1^T + \cdots + v_k v_k^T) x_j = \sum_{i=1}^{k} v_i^T \frac{1}{n}\sum_{j=1}^{n} (x_j x_j^T) v_i$$

$$= \sum_{i=1}^{k} v_i^T \, S v_i = \sum_{i=1}^{k} v_i^T \, \lambda_i v_i = \sum_{i=1}^{k} \lambda_i \qquad (2.3.10)$$

Hence, the variance explained can be written as a ratio over sums over eigenvalues of the covariance matrix **S**. ■

**Theorem 2.3.4:** (Alternative Interpretation of PCA)

*An alternative interpretation of PCA is that it finds the major axis of variation in the dataset such that:*

i. *The first PC defines the direction in the dataset with the greatest variance.*
ii. *The second PC defines a direction that*
    a) *is orthogonal to the first PC, and*
    b) *represents the major direction of the remaining variance in the dataset.*
iii. *In general, the $i^{th}$ PC is orthogonal to all previous i-1 PCs and represents the direction of maximum variance remaining in the dataset.*

***Proof: (i) Variance maximization along first PC***

We start by trying to find one PC, $v_1$, that maximizes the variance of *X*:

$$maximize_{\ v_1} \, Var(v_1^T X) = maximize_{v_1} \ v_1^T \mathbf{S} v_1 \text{ , with } v_1^T v_1 = 1.$$

Then we form the Lagrangian to solve this problem as:

$$v_1^T \mathbf{S} v_1 - \lambda(v_1^T v_1 - 1) \qquad (2.3.11)$$

and take the derivative with respect to $v_1$ and set it to zero:

$$\frac{\partial}{\partial v_1}(v_1^T \mathbf{S} v_1 - \lambda(v_1^T v_1 - 1)) = 0$$

$$\Rightarrow \frac{\partial}{\partial v_1}(v_1^T \mathbf{S} v_1) - \lambda \frac{\partial}{\partial v_1}(v_1^T v_1 - 1) = 2\mathbf{S} v_1 - 2\lambda v_1 = 0$$

$$\Rightarrow \mathbf{S} v_1 = \lambda v_1$$

Thus, the solution $v_1$ is an eigenvector of the matrix **S**, and then multiplying the above equation from the left by $v_1^T$ gives $v_1^T \mathbf{S} v_1 = v_1^T \lambda v_1 = \lambda v_1^T v_1 = \lambda = \lambda_1$.

So, the variance is maximized by picking the eigenvector corresponding to the largest eigenvalue. Hence, the first eigenvector $v_1$ is the direction of the principal component (PC) that maximizes the variance of $Xv_1$.

***(ii) Variance maximization along the second PC***

The second direction of projection should be independent from the first one,

$Cov(v_2^T X, v_1^T X) = 0$. This can be written as:

$Cov(v_2^T X, v_1^T X) = v_2^T X X^T v_1 = v_2^T \mathbf{S} v_1 = v_2^T \lambda v_1 = \lambda v_2^T v_1 = 0$, which implies $v_2^T v_1 = 0$, *as* $\lambda \neq 0$.

Here we will try to find $v_2$ such that:

$$maximize_{\ v_2^T v_2 = 1, v_2^T v_1 = 0}\ Var(v_2^T X) = maximize_{v_2^T v_2 = 1, v_2^T v_1 = 0}\ v_2^T \mathbf{S} v_2$$

Then we form the Lagrangian as: $v_2^T \mathbf{S} v_2 - \lambda(v_2^T v_2 - 1) - \mu(v_2^T v_1)$, and set the derivative with respect to $v_2$ *to zero:*

$$\frac{\partial}{\partial v_2}(v_2^T \mathbf{S} v_2 - \lambda(v_2^T v_2 - 1) - \mu(v_2^T v_1)) = 2\mathbf{S} v_2 - 2\lambda v_2 - \mu v_1 = 0.$$

If we multiply this from the left by $v_1^T$, we get:

$$2v_1^T \mathbf{S} v_2 - 2v_1^T \lambda v_2 - v_1^T \mu v_1 = -v_1^T \mu v_1 = -\mu = 0.$$

Now, $\mu = 0$ implies that $\boldsymbol{S} v_2 = \lambda v_2$, showing that $v_2$ is again an eigenvector of matrix **S**, and we again pick the eigenvector corresponding to the second largest eigenvalue to maximize the variance along the second PC. The proofs for the other PC (for k > 2) follow the same scheme. ■

***Algorithm 2.1:*** *PCA dimensionality reduction using Covariance matrix*

**Input:** A matrix $X \in \mathbb{R}^{d \times n}$; number of components *k;*

1. for $i = 1,2, \dots, n$ set $x_i \longleftarrow x_i - \frac{1}{n}\sum_{i=1}^{n} x_i$
2. if *n* > *d* then
3. $\boldsymbol{S} = \frac{1}{n} X X^T$
4. Compute the r leading eigenvectors $v_1, \dots, v_k$ *of* $\boldsymbol{S}$.
5. else
6. $K = \frac{1}{n} X^T X$
7. Compute the *k* leading eigenvectors $v_1^*, \dots, v_k^*$ *o*f K.
8. for $i = 1,2, \dots, k$ set $v_i = \frac{1}{\|X v_i^*\|} X v_i^*$.
9. end if
10. return reduction matrix $V = [v_1, \dots, v_k]^T$ or reduced points *VX*

# CHAPTER

# 3 LITERATURE REVIEW

This section aims to review the published related work in the past recent years that used features dimensionality reduction and classification approaches on high-dimensional datasets.

## 3.1 Green Coffee Beans

Mehari, B. et al. (2016) investigated the possibility of using the phenolic profiles of green coffee beans, together with statistical pattern recognition techniques to identify characteristic marker compounds to distinguish Ethiopian coffees according to their region of origin. The study used PCA and identified the concentrations of 3-O-caffeoylquinic and 4,5-O-dicaffeoylquinic acids as the characteristic markers for Northwest and East (Harar) regions coffees, respectively. Moreover, they applied LDA model for the classification of the green coffee beans and achieved the recognition and prediction abilities of 91% and 90%, respectively, at regional level, and 89% and 86%, respectively, at sub-regional level.

F Kurniawan et al., (2019) proposed the NIR spectroscopy for characterization and classification of intact Java arabica coffee beans based on their origin. Three kinds of Java arabica coffee beans namely Arabica Java Preanger, Arabica Bondowoso and Arabica Malang were used in their research. Discriminant analysis (DA) of Principle Components was developed to classify coffee beans based on their origin. The results showed that PC analysis using PC1 and PC2 gave the best results for discriminating three kinds of coffee beans. The DA of three principle components of reflectance data could classify Arabica coffee beans accurately (100%).

## 3.2 Data Normalization

Data standardization or normalization is a data preprocessing technique to transform the given data into a desired range, which improves the data quality and makes them fit for use. Data is usually standardized in order to remove inconsistency and create the desirable data pattern or structure. Since the datasets are taken from the source that contains some noise and redundant data, it is required to do normalization before dimensionality reduction and classification procedures.

Zhang et al. (2003) discussed the need for data normalization or scaling techniques. Due to some disturbances like noise (ambiguity), data collected and kept in the storage may provide low quality information. So, it is suggested to clean the data before it is used for dimension reduction and classification as well as prediction. Kotsiantis et al. (2006) investigated the importance of data normalization techniques in learning algorithm, stating that the original data can be formatted and standardized clearly using the appropriate data normalization techniques.

Ting Li, et al. (2017) investigated Adaptive Scaling, which have great overall performance and showed that it was much better than z-Score Standardization and Rang Scaling (Min-Max), the two widely used methods. They also generalized Adaptive Scaling and fit it to high dimensional data. They discussed that Tree model does a good job and is scale-invariant. However, Adaptive Scaling applied on high dimensional data, followed by Neural Network, has the highest accuracy (82.15%) (Ting Li, et al., 2017).

Jayalakshmi and Santhakumaran (2009) investigated the statistical normalization techniques like z-score, min-max and median. They presented that those standardization techniques enhance the accuracy as well as reliability of a classifier model. Rathod and Momin (2012) evaluated the performance of standardization (normalization) in outlier detection. They investigated that to refine the results from the given dataset, normalization techniques such as z-score, min-max and decimal scaling were implemented in outlier detection from those data. To avoid the missing and redundant values in gene expression array data, Chandrasekhar et al. (2011) studied the different data normalization and clustering techniques.

## 3.3 Dimensionality Reduction

The study by Xia, Yang and Li (2010) used Principal Component Analysis (PCA) as the dimension reduction techniques and Grey Neural Networks as the classifier on the ‘KDD-99’ dataset to implement an intrusion detection system. Moreover, in the same study (Xia, Yang and Li, 2010), they used PCA for dimensions reduction while Decision tree and Nearest Neighbor as classifiers on ‘KDD-99’. Similarly, Vasan and Surendiran (2016) studied the efficiency of PCA for intrusion detection, and they fulfilled their experiments using Random Forest and C4.5 on KDD-CUP (Bay et al., 2000) and UNB-ISCX (Shiravi, A. et al., 2012).

Pinderjeet Kaur (2012) used Principal Component Analysis (PCA), which allowed them to minimize computational cost of Content Based Image Retrieval (CBIR). Arunasakthi and Kamatchipriya (2014) conducted review on linear and non-linear dimensionality reduction techniques, and stated that Principal Component Analysis and Linear Discriminant Analysis were the fundamental techniques for dimensionality reduction as well as retrieving effective variables or factors of high dimensional data points in the given dataset. The investigation on dimensionality reduction techniques by Julie M. David and Kannan Balakrishanan (2014) showed that principal component analysis gave a set of new variables, which involved data standardization, orthonormal vectors' computation by PCA and sorting the principal components in descending order of their weight.

In their comparative study of PCA, ICA, and LDA on the "FERET Data Set", Kresimir Delac, Mislav Grgic and Sonja Grgic (2006), found PCA as effective dimensional reduction technique and achieved in retaining the most suitable components in which the projection of sample does maintain key information of the original dataset. The review "CBIR Feature Vector Dimension Reduction with Eigenvectors of Covariance Matrix using Row, Column and Diagonal Mean Sequences", Dr. H.B. Kekre, Sudeep D. Thepade and Akshay Maloo (2010) noted PCA as a transformational that converts each image to its corresponding eigen-image in the database.

Dimensionality reduction techniques that improve the performance of support vector machines in classification techniques were discussed by Gupta et al. (2002). They compared the linear and non-linear (kernel) methods for face (image) recognition. They discussed that techniques such as Principal Component Analysis, Kernel-based Principal Component Analysis, Linear Discriminant Analysis and Kernel-based Linear Discriminant Analysis were applied to reduce the dimensionality of a given dataset while the reduced dataset were classified using support vector machine and nearest neighbor. They concluded that together principal component analysis and nonlinear support vector machine gave better result.

Omucheni et al. (2014) applied principal component analysis to perform dimensionality reduction and to enhance PCA score images for better visualization. They used PCA as a feature extraction through clusters in PCA score space. The study conducted by Nandi, D. A. (2015) investigated the applications of PCA in identifying various medical images and the results obtained from those images showed better efficacy of the techniques. Nazlibilek et al.,

(2015) used the principal component analysis method for classification of five types of white blood cells and revealed effectiveness of PCA in their study.

## 3.4 Classification

Classification technique is based on the inductive learning principle that finds and separate patterns from the given dataset. Limère et al. (2004) developed a model with decision tree induction principle applied for firm growth. They achieved interesting results and fit the model to economic data such as growth ambitions and growth potential, and also to growth competence and resources. Hoi et al. (2006) introduced framework of learning the unified kernel-based machines used for both unlabeled and labeled data. Their framework of learning includes supervised learning, semi-supervised learning and active learning. Moreover, to classify the given unlabeled and labeled data efficiently, they proposed and implemented a spectral kernel.

Xu et al. (2008) proposed a reproducing kernel Hilbert space for information theoretic learning. The framework used the symmetric nonnegative definite kernel-function. Though this framework gave better result than the previous "RKHS" frameworks, still there is difficulty in selecting an appropriate kernel function for a particular area. Shilton and Palaniswami (2008) proposed a unified approach to support vector machines, which was formulated for binary classification and later extended to regression model and one-class classification. A binary classification framework for two stage-multiple kernel learning was proposed by Kumar et al. (2012), which is easier to leverage research in binary classification and to develop robust and scalable kernel-based techniques.

Takeda et al. (2012) developed a unified robust classification model that enhances and optimizes the existing classification models such as linear discriminant analysis and support vector machine, which provides several benefits like extending the existing techniques, providing well-defined theoretical results and describes the relationships among existing models.

# CHAPTER 4 RESEARCH METHODS

This chapter describes the research methodology and the significant steps that have been involved. An outline of the research framework is given in Figure 4.1. The proposed dimensionality reduction and classification framework can classify the green coffee dataset and achieve better results. From the literature, it is found that feasible empirical models were constructed using PCA and LDA. The steps involved in the dimensionality reduction and classification framework are listed below.

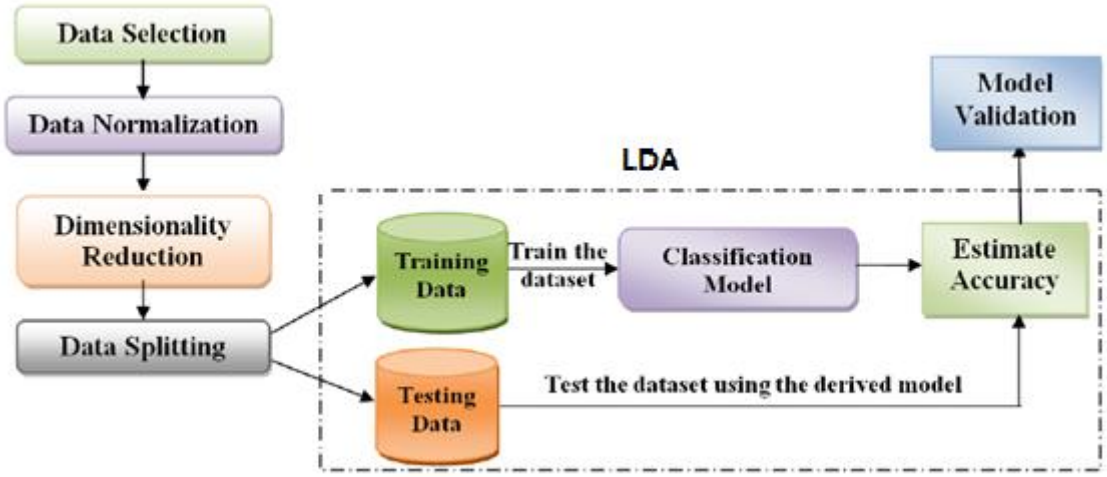



**Figure 4.1 Proposed DR and Classification Framework**

## 4.1 Dataset

This study applied PCA and LDA on the samples of green coffee beans to identify the most discriminating compounds and classify the sample coffees based on the geographical origin. The data source and methods of data analysis are presented as follows.

### 4.1.1 Data Source

The source of the data for this study was the study conducted by (Mehari, B. et al., 2016), "*Journal of Food Composition and Analysis* 45 (2016) 16–2518". The data were constructed by applying the Box-Muller method on the mean and standard deviation values given for each regional and sub-regional category at (Mehari, B. et al., 2016). Thus, the data were

organized in such a way while considering the minimum and maximum values for each category (i.e. both at regional and sub-regional levels).

The Box–Muller transform was developed as a more computationally efficient alternative to the statistical inverse transform sampling method (Kloeden and Plate, 1992). Studies show that the Box–Muller transform is efficient and superior for processors (e.g. GPUs or modern CPUs) with vector units (Howes, Lee, 2008). Furthermore, the Box–Muller transform employed for drawing resulted from truncated bivariate of Gaussian densities (Martino, L. et al., 2012).

The coffee samples were collected from the four sampling regions, which are the major coffee production areas across Ethiopia, such as East, Northwest, West and South categories (Mehari, B. et al., 2016). Accordingly, the dataset contained a total of 100 samples, 27 from East, six from Northwest, 18 from West and 49 from South, and within each of these regional categories (with the exception of East), different sub-regional sampling areas were included. The geographical regions of origin and varieties (Mehari, B. et al., 2016) of the samples of green coffee bean were given as in Table 4.1.

**Table 4.1: Geographical regions of origin and varieties of the green coffee bean samples**

| Region | Sub-regional coffee type | Number of samples | Sample ID[a] |
|---|---|---|---|
| East | Harar | 27 | H1–H27 |
| West | Jimma A | 3 | JA1–JA3 |
| | Jimma B | 3 | JB1–JB3 |
| | Kaffa | 10 | K1–K10 |
| | Wollega | 2 | W1–W2 |
| Northwest | Benishangul | 3 | B1–B3 |
| | Finoteselam | 3 | F1–F3 |
| South | Sidama SA | 10 | SA1–SA10 |
| | Sidama SB | 29 | SB1–SB29 |
| | Yirgachefe | 10 | Y1–Y10 |

[a] Sample identification used in this study

Source: Mehari, B. et al. (2016)

The dataset consisted of 10 variables and 100 observations, corresponding to the samples of the green coffee beans. Among these ten variables, eight of them were found to be compound contained in green coffee beans, and two of them are their concentrations ratios. The phenolic fractions of green coffee beans, produced in Ethiopia, were found to contain predominantly Chlorogenic acids (CGAs), as is typical compound for the green coffees (Mehari, B. et al., 2016). The compounds and concentration rations, which were the variables of this study, were presented in Table 4.2.

**Table 4.2: The ten variables of the dataset used in this study**

| ID[b] | Variable Name | Description |
|---|---|---|
| qa3 | 3-CQA | 3-O-caffeoylquinic acid |
| qa5 | 5-CQA | 5-O-caffeoylquinic acid |
| qa4 | 4-CQA | 4-O-caffeoylquinic acid |
| qa5p | 5-pCoQA | 5-O-p-coumaroylquinic acid |
| qa5f | 5-FQA | 5-O-feruloylquinic acid |
| qa34 | 3,4-diCQA | 3,4-di-O-caffeoylquinic acid |
| qa35 | 3,5-diCQA | 3,5-di-O-caffeoylquinic acid |
| qa45 | 4,5-diCQA | 4,5-di-O-caffeoylquinic acid |
| qa35to45 | 3,5-diCQA/4,5-diCQA | 3,5-di-O-caffeoylquinic acid to 4,5-di-O-caffeoylquinic acid concentration ratio |
| qa45to34 | 4,5-diCQA/3,4-diCQA | 4,5-di-O-caffeoylquinic to 3,4-di-O-caffeoylquinic acid concentration ratio |

[b] Variable identification used in this study

### 4.1.2 Methods of Data Analysis

In this work, the statistical package for social science (SPSS) and MATLAB software are used to analysis the data. First, prior to constructing PCA model, the suitability of dataset was assessed. Kaiser-Meyer-Olkin (KMO) and Bartlett's measures statistics were used to assess suitability of the dataset for basis of PCA (Maat, Zakaria, Nordin, & Meerah, 2011). KMO measure of sampling adequacy and Bartlett's test of sphericity were conducted on the green coffee beans dataset. The ten variables of our dataset were subjected to principal component

analysis (PCA). The sampling is adequate if the value of KMO test value is greater than 0.5 (Field, 2000; Kaiser, 1974), and the Bartlett's Test of Sphericity (BTS) must be significant at $p < .05$ (Hair et al., 2010; Pallant, 2007; Tabachnick and Fidell, 2007).

Second, data normalization was applied to transform the raw data into a standard form so as to ease the algorithm's process, and then data were analyzed using the multivariate statistical techniques of principal component analysis (PCA) and one-way analysis of variance (ANOVA). One-way ANOVA was used to test for the presence of significant differences between the mean concentrations of the compounds in the green coffee beans both from different regional and sub-regional categories. Differences were considered significant when $p < 0.05$. The loadings plots, the Score plots, and Superposition plots from PCA were used to identify the variables (compound) and the corresponding coffee samples. The Superposition plots and the significant differences revealed by ANOVA were used to select the suitable discriminant markers for the corresponding coffee samples. Finally, discriminant analysis was applied to develop LDA classification models that can be used to classify the samples and predict the geographical origin of green coffee beans.

## 4.2 Data Normalization

Data normalization is the main preprocessing techniques that prepare the data before the construction of dimensionality reduction and classifier models. Data normalization is used to transform the raw data into a standard form, which contributes in simplifying the process of the algorithms. Due to standardization (or normalization) process, the rationale of the given dataset can be changed. Hence, the proposed techniques or algorithms should be checked with the standardized dataset to keep the information content untouched.

Data normalization is an important data transformation method in order to improve the accuracy and achieve better performance in given dataset. Normally, the values of the variables in the dataset are measured in different scales such that some features may be decimals while others may be integer values. So that aim of applying data normalization to a given dataset is to organize and manage the values of the variables in the dataset. Moreover, normalization scales the values of the variables (features) to the same range. Since the input data should not be submerged by other data values in terms of distance metric, so that a preprocess data normalization is used in dimensionality reduction and classification

techniques. Each feature (variable) value starts in the same range, which minimizes bias and speeds up the dimensionality reduction and classification process.

The popular Z-score normalization (Kotsiantis et al. 2006), which is known as zero-mean standardization, normalizes the values of the data points in the dataset using mean and standard deviation. The mean and standard deviation for each variable (or column) vector is calculated across the given dataset. This normalization helps in determining whether a value of the data point is below or above the average value. It is recommended to normalize the original dataset when the maximum or minimum values of the attribute are unknown and outliers dominate the values of the data points.

The Z-score technique transforms a value v to v' by:

$$v' = \frac{v - \bar{v}}{\sigma}$$

where $v'$ is a new value of an attribute , v is an old value of an attribute, $\bar{v}$ is the mean of an attribute value v and $\sigma$ is the standard deviation of an attribute value v.

## 4.3 Principal Component Analysis (PCA) and its Algorithm

PCA, dimensionality reduction technique, aims to enhance the performance of the data reduction and classification models by removing the irrelevant data within the dataset. Reduction of the number of attributes in the dimensionality reduction as well as classification models helps to alleviate the problem known as the curse of dimensionality, which is the major crisis to data storage and retrieval (Larose, 2005).

Principal component analysis (PCA), which is a popular multivariate method in data analysis, computes new uncorrelated linear combinations of the original variables (features) so that the first linear combination captures the largest variance; the second linear combination explains the second largest variance; and so on. These new computed linear combinations are called the principal components (PCs) of the PCA model.

After the principal components (PCs) are computed, data analysis, visualization and interpretation can then be performed using those PCs instead of the original variables of the dataset. Most textbooks in advanced multivariate techniques include a chapter and large coverage regarding PCA, to mention some, Mardia et al. (1979), Johnson and Wichern

(2002) are among the major ones. Specially, Jolliffe (2002) published a book dedicated to principal component analysis that used as a major source for this work. Hence, the PCA technique is computed as follows.

### 4.3.1 Principal Component Analysis techniques

PCA can employ the orthogonal projection to convert a large dataset of possibly interrelated variables into a smaller set of linearly uncorrelated variables (or indicators) (Doorsamy and Cronje, 2015; Jolliffe, I.T., 2002).

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n vectors of the dataset $\boldsymbol{X}$ in $\mathbb{R}^d$. We assume that they are mean-centered. The goal of PCA is to reduce the dimensionality of these vectors using a linear transformation. The matrix $V \in \mathbb{R}^{k \times d}$, where $k < d$, induces a mapping $\boldsymbol{x} \rightarrow \boldsymbol{V}\boldsymbol{x}$, where $V\boldsymbol{x}$ is the lower dimensional representation of $\boldsymbol{x}$. A second matrix $U \in \mathbb{R}^{d \times k}$can be used to approximately recover each original vector $\boldsymbol{x}$ from its compressed (reduced) form.

The interpretation of PCA is that it finds the major axis of variation in the dataset such that the first PC defines the direction in the dataset with the greatest variance.

Let $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T$ be the covariance matrix of the data matrix $\boldsymbol{X}$ with mean-centered.

Now, we can find one PC ($v_1$) that maximizes the variance of $X$:

$$maximize\ _{v_1}\ Var(v_1^T X) = maximize_{v_1}\ v_1^T \mathbf{S} v_1, \text{ with } v_1^T v_1 = 1, \textit{(See section 2.3).}$$

Then, we form the Lagrangian to solve this problem as: $v_1^T \mathbf{S} v_1 - \lambda(v_1^T v_1 - 1)$

and take the derivative with respect to $v_1$ and set it to zero:

$$\frac{\partial}{\partial v_1}(v_1^T \mathbf{S} v_1 - \lambda(v_1^T v_1 - 1)) = 0$$

$$\Rightarrow \frac{\partial}{\partial v_1}(v_1^T \mathbf{S} v_1) - \lambda \frac{\partial}{\partial v_1}(v_1^T v_1 - 1) = 2\mathbf{S}v_1 - 2\lambda v_1 = 0 \Rightarrow \mathbf{S}v_1 = \lambda v_1$$

Thus, the solution $v_1$ is an eigenvector of the matrix $\mathbf{S}$, and then multiplying the above equation from the left by $v_1^T$ gives $v_1^T \mathbf{S} v_1 = v_1^T \lambda v_1 = \lambda v_1^T v_1 = \lambda = \lambda_1$.

So, the variance is maximized by picking the eigenvector corresponding to the largest eigenvalue. Hence, the first eigenvector $v_1$ is the direction of the principal component (PC) that maximizes the variance of $Xv_1$.

The second PC defines a direction that is orthogonal to the first PC, and it represents the major direction of the remaining variance in the dataset. The second direction of projection should be independent from the first one, $Cov(v_2^T X, v_1^T X) = 0$.
This can be written as:

$Cov(v_2^T X, v_1^T X) = v_2^T X X^T v_1 = v_2^T \mathbf{S} v_1 = v_2^T \lambda v_1 = \lambda v_2^T v_1 = 0$, which implies $v_2^T v_1 = 0, as\ \lambda \neq 0.$

Here we will try to find $v_2$ such that:

$$maximize_{\ v_2^T v_2 = 1, v_2^T v_1 = 0}\ Var(v_2^T X) = maximize_{v_2^T v_2 = 1, v_2^T v_1 = 0}\ v_2^T \mathbf{S} v_2$$

Then we form the Lagrangian as: $v_2^T \mathbf{S} v_2 - \lambda(v_2^T v_2 - 1) - \mu(v_2^T v_1)$, and set the derivative with respect to $v_2$ *to zero:*

$$\frac{\partial}{\partial v_2}(v_2^T \mathbf{S} v_2 - \lambda(v_2^T v_2 - 1) - \mu(v_2^T v_1)) = 2\mathbf{S} v_2 - 2\lambda v_2 - \mu v_1 = 0.$$

If we multiply this from the left by $v_1^T$, we get:

$$2v_1^T \mathbf{S} v_2 - 2v_1^T \lambda v_2 - v_1^T \mu v_1 = -v_1^T \mu v_1 = -\mu = 0.$$

Now, $\mu = 0$ implies that $\boldsymbol{S} v_2 = \lambda v_2$, showing that $v_2$ is again an eigenvector of matrix **S**, and we again pick the eigenvector corresponding to the second largest eigenvalue to maximize the variance along the second PC. In general, the i[th] PC is orthogonal to all previous i-1 PCs and represents the direction of maximum variance remaining in the dataset.

***Algorithm 4.1:*** **PCA techniques and algorithm**

**Step 1:** *Standardize the original dataset*. PCA is commonly applied to a dataset consisting of *n* samples (observations) with *d*-features (dimensions), thus the original dataset can be expressed as a matrix $\boldsymbol{X}_{n \times d}$. Due to the different measured scales among the variables of coffee beans, scaling (standardization) is applied to the dataset to enable good comparability between variables. Accordingly, the original data matrix $\boldsymbol{X}_{n \times d}$ can be transformed into a standardized matrix $\boldsymbol{Y}_{n \times d}$ with zero mean and unit variance as shown in (4.3.1) below.

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}} \qquad (4.3.1)$$

where i = 1, 2, … , n, j = 1, 2, … ,d, $x_{ij}$ is the j[th] variable value of the i[th] sample in the original data matrix $\boldsymbol{X}_{n\times d}$, $\mathbf{x}_j$ is the j[th] variable of $\boldsymbol{X}_{n\times d}$, , $y_{ij}$ is the standardized value of $x_{ij}$, while $\bar{x}_j$ and $\sigma_{x_j}$ are the mean and standard deviation of $\mathbf{x}_j$, respectively.

**Step 2:** *Calculate the covariance matrix* **S** using the equation:

$$\mathbf{S} = (s_{y_i,y_j})_{d\times d} = \frac{1}{n-1} Y^T Y \qquad (4.3.2)$$

where $y_i$ and $y_j$ are the i[th] and j[th] column vector of $\boldsymbol{Y}_{n\times d}$, respectively, $s_{y_i,y_j}$ is the covariance value between $y_i$ and $y_j$, which is known as correlation coefficient between the i[th] and j[th] variables.

**Step 3:** *Compute eigenvalues and eigenvectors of* **S***:* They are computed using $|\mathbf{S} - \lambda I| = 0$. Then eigenvalues, which arranged in descending order, are $\lambda_1 > \lambda_2 >, \dots, > \lambda_d$. There is a corresponding eigenvector for each eigenvalue. Using the equation $\mathbf{S}v_j = \lambda_j v_j$, $v_j$ (the unit eigenvector corresponding to $\lambda_j$) is calculated and $\sum_{i=1}^{d} v_{ij}^2 = 1$. Therefore, $\boldsymbol{V} = (v_1, v_2, \dots, v_d)$ is a unit orthogonal matrix consisting of all the unit eigenvectors obtained.

**Step 4:** *Determine principal components*. The criterion: eigenvalue greater than one or total cumulative percentage variance greater than 80% is used. The cumulative percentage variance is determined by using $\alpha_j = \lambda_j / \sum_{j=1}^{d} \lambda_j$ and $\beta_k = \sum_{i=1}^{k} \lambda_i / \sum_{j=1}^{d} \lambda_j$ where $\alpha_j$ is percentage variance of i[th] PC, $\beta_k$ is total cumulative percentage variance of $k$ PCs (with $k \leq d$). When $\beta_k$ greater than 80% is achieved, $k$ PCs are selected for further process.

**Step 5:** *Identify the variables belonging to those determined PCs*. The component loading of each variable on each those determined PC, namely, correlation coefficient $\theta_{ij}$ between the i[th] indicator and j[th] PC is computed by using $\theta_{ij} = v_{ij}\sqrt{\lambda_j}$, where $\lambda_j$ is the eigenvalue corresponding to j[th] PC, $v_{ij}$ is i[h] value of $\boldsymbol{v_j}$. Only the component loadings with $|\theta_{ij}| \geq 0.5$ is considered for each determined PC.

**Step 6:** *Calculate component (PC) scores*. PCA scores (**f**1, **f**2, … , **f**p) can be computed by using the projection **F** = **YV**. Besides, the percentages of variation explained by each

PC are used as weights, which is calculated by $\omega_k = \frac{\lambda_k}{\sum_{k=1}^{p} \lambda_k}$. Therefore, the overall PCA score ***CF*** is computed using Equation (4.3.3) below.

$$CF = \sum_{k=1}^{p} \omega_k f_k \qquad (4.3.3)$$

### 4.3.2 Improved PCA techniques

Equation (4.3.1) at the beginning of above algorithm makes the variance of each variable equal to 1, which reduces the influence of the spread of data (or dispersion degree differences) on PCs (Shang and Wang, 2014). Thus, PCs computed from the normalized (standardized) dataset could not fully reflect information of the original dataset (Hosseini and Kaneko, 2011; Cai, W. et al., 2016). Based on this fact, an improved normalization method is proposed, that aimed to improve the spread (dispersion) of data points around the mean, as shown below in (4.3.4).

$$z_{ij} = (x_{ij} - \bar{x_j})/\beta_{x_j} \qquad (4.3.4)$$

Where $z_{ij}$ is the standardized value of $x_{ij}$, $\beta_{x_j} = \max(x_j) - \min(x_j)$ and $\beta_{x_j} > 0$

The mean and standard deviation of j[th] indicator ($z_j$) in standardized data matrix $\boldsymbol{Z}_{n\times d}$ can be computed using Equations (4.3.5) and (4.3.6), respectively.

$$\bar{z_j} = \sum_{i=1}^{n} \frac{z_{ij}}{n} = \sum_{i=1}^{n} \frac{\frac{x_{ij}-\bar{x}_j}{\beta_{x_j}}}{n} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)}{n\beta_{x_j}} = 0 \qquad (4.3.5)$$

$$\sigma_{z_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_{ij} - \bar{z_j})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_{ij}-\bar{x}_j}{\beta_{x_j}})^2} = \frac{\sigma_{x_j}}{\beta_{x_j}} \qquad (4.3.6)$$

$$\rho_{z_j,z_k} = \frac{C_{z_j,z_k}}{\sigma_{z_j}\sigma_{z_k}} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{ij} - \bar{z_j})(z_{ik} - \bar{z_k})/(\sigma_{z_j}\sigma_{z_k})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ij} - \bar{x}_j)}{\beta_{x_j}}\frac{(x_{ik} - \bar{x}_k)}{\beta_{x_k}}/(\frac{\sigma_{x_j}\sigma_{x_k}}{\beta_{x_j}\beta_{x_k}})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/(\sigma_{x_j}\sigma_{x_k})$$

$$= \rho_{x_j,x_k} \qquad (4.3.7)$$

According to Equation (4.3.7), $\boldsymbol{Z}_{n\times d}$ and $\boldsymbol{X}_{n\times d}$ have the same correlation coefficient matrix, indicating that the improved standardization method keeps correlation information of all indicators. Importantly, dispersion degree differences of all indicators are partly retained according to Equation (4.3.6), and the classical PCA is to some extent improved.

***Algorithm 4.2:*** **Improved PCA algorithm**

**1.** *Standardize the original dataset*. The original data matrix $\boldsymbol{X}_{n\times d}$ can be transformed into a standardized matrix $\boldsymbol{Z}_{n\times d}$ as:

$$z_{ij} = (x_{ij} - \bar{x_j})/\beta_{x_j}$$

Where i = 1, … , n, j = 1, … ,d, $x_{ij}$ is the j[th] variable value of the i[th] sample in $\boldsymbol{X}_{n\times d}$, $\bar{x}_j$ is the mean of $\mathbf{x}_{\text{j,,}}$ $z_{ij}$ is the standardized value of $x_{ij}$, $\beta_{x_j} = \max(x_j) - \min(x_j)$, $\beta_{x_j} > 0$

**2.** *Calculate the covariance matrix* **S** using the equation:

$$\mathbf{S} = (s_{z_i,z_j})_{d\times d} = \frac{1}{n-1}\boldsymbol{Z}^T\boldsymbol{Z}$$

where $z_i$ and $z_j$ are the i[th] and j[th] column vector of $\boldsymbol{Z}_{n\times d}$, respectively, $s_{y_i,y_j}$ is the covariance value between $z_i$ and $z_j$.

**3.** *Compute eigenvalues and eigenvectors of* **S*:*** They are computed using$|\mathbf{S} - \lambda I| = 0$. Then eigenvalues, which arranged in descending order, are $\lambda_1 > \lambda_2 >, \dots, > \lambda_d$. There is a corresponding eigenvector for each eigenvalue calculated using $\mathbf{S}v_j = \lambda_j v_j$ and organized as columns of $\boldsymbol{V} = (v_1, v_2, \dots, v_d)$.

**4.** *Determine principal components*. The criterion: eigenvalue greater than one or total cumulative percentage variance greater than 80% is used.

**5.** *Identify the variables belonging to those determined PCs*. The component loading of each variable on each those determined PC is computed by using $\theta_{ij} = v_{ij}\sqrt{\lambda_j}$, where $\lambda_j$ is the eigenvalue corresponding to j[th] PC, $v_{ij}$ is i[h] value of $\boldsymbol{v_j}$.

**6.** *Calculate component (PC) scores*. PCA score is computed using the projection **F** = **ZV**. The overall PCA score $\boldsymbol{CF}$ is computed using: $CF = \sum_{k=1}^{p}\omega_k f_k$, where $\omega_k = \frac{\lambda_k}{\sum_{k=1}^{p}\lambda_k}$ is the percentages of variation explained by each PC.

## 4.4 LDA Techniques and its Algorithm

The goal of the LDA technique is to project the original data matrix onto a lower dimensional space. To achieve this goal, three steps needed to be performed.

The first step is to calculate the distance between the means of different classes, which is called the between-class variance or between-class matrix, and followed by computing the distance between the mean and the data points of each class, which is called the within-class variance or within-class matrix. Finally, we construct the LDA lower dimensional space, which simultaneously maximizes between-class and minimizes within-class variances. Next, these three steps are presented in detail, and followed by the LDA algorithm.

**Step 1: Computing the Between-Class Variance ($S_B$)**

The between-class variance of the $i^{th}$ class ($S_{Bi}$ ) represents the distance between the mean of the $i^{th}$ class ($\mu_i$) and the total mean ($\mu$). LDA technique searches for a lower-dimensional space, which simply maximize the separation distance between classes. To explain how the between-class variance or the between-class matrix ($S_B$) can be computed, the following assumptions are made. Given the original data matrix $\boldsymbol{X} = \{x_1, x_2, \dots, x_n\}$, where $x_i$ represents the $i^{th}$ sample, pattern, or observation and n is the total number of samples. Each sample is represented by d features ($x_i \in \mathbf{R}^d$). In other words, each sample is represented as a point in d-dimensional space.

Assume the data matrix is partitioned into c classes. The total number of samples (n) is computed using $n = \sum_{i=1}^{C} n_i$, where $n_i$ represents the number of samples of the $i^{th}$ class.

To calculate the between-class variance ($S_B$), the separation distance between different classes, which is denoted by $(m_i - m)$ is calculated using:

$$(m_i - m)^2 = (W^T\mu_i - W^T\mu)^2 = \mathrm{W}^T(\mu_i - \mu)(\mu_i - \mu)^T W \qquad (4.4.1)$$

where $m_i$ represents the projection of the mean of the $i^{th}$ class and it is calculated as follows, $m_i = W^T\mu_i$, where m is the projection of the total mean of all classes and it is calculated as follows, $m = W^T\mu$, W represents the transformation matrix of LDA, $\mu_{i,}(1xd)$- dimensional, represents the mean of the $i^{th}$ class and it is computed as in Equation (4.4.2), and $\mu,(1xd)$-dimensional, is the total mean of all classes and it can be computed as in Equation (4.4.3).

$$\mu_j = \frac{1}{n_j}\sum_{x_i \in \omega_j} x_i \quad (4.4.2)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{C} \frac{n_i}{n}\mu_i \quad (4.4.3)$$

where c represents the total number of classes.

The term $(\mu_i - \mu)(\mu_i - \mu)^T$ in Equation (4.4.1) represents the separation distance between the mean of the i$^{th}$ class ($\mu_i$) and the total mean ($\mu$), or simply it represents the between-class variance of the i$^{th}$ class ($S_{Bi}$ ).

Substitute $S_{Bi}$ into Equation (1) as follows:

$$(m_i - m)^2 = W^T S_{B_i} W \quad (4.4.4)$$

Therefore, the total between-class variance is calculated as follows,

$$S_B = \sum_{i=1}^{C} n_i \, S_{B_i}$$

**Step 2: Computing the Within-Class Variance ($S_W$)**

The within-class variance of the ith class ($S_{Wi}$ ) represents the difference between the mean and the samples of that class. LDA technique searches for a lower-dimensional space, which is used to minimize the difference between the projected mean ($m_i$) and the projected samples of each class ($W^T x_i$), or simply minimizes the within-class variance (Yu and Lu et al., 2003). The within-class variance of each class ($S_{Wj}$) is calculated as in Equation (4.4.5).

$$\sum_{x_i \in \omega_j, j=1,\dots,C} \left(W^T x_i - m_j\right)^2 = \sum_{x_i \in \omega_j, j=1,\dots,C} \left(W^T x_{ij} - W^T \mu_j\right)^2$$

$$= \sum_{x_i \in \omega_j, j=1,\dots,C} W^T (x_{ij} - \mu_j)^2 W$$

$$= \sum_{x_i \in \omega_j, j=1,\dots,C} W^T \left(x_{ij} - \mu_j\right)(x_{ij} - \mu_j)^T W$$

$$= \sum_{x_i \in \omega_j, j=1,\dots,C} W^T S_{W_j} W \quad (4.4.5)$$

From Equation (4.4.5), the within-class variance for each class can be reformulated as:

$$S_{W_j} = d_j^T * d_j = \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$$

where $x_{ij}$ represents the $i^{th}$ sample in the $j^{th}$ class, and $d_j$ is the centering data of the $j^{th}$ class, i.e. $d_j = \omega_j - \mu_j = \{x_i\}_{i=1}^{n_j} - \mu_j$.

The total within-class variance represents the sum of all within-class matrices of all classes and it can be calculated as in Equation (4.4.6).

$$S_W = \sum_{i=1}^{C} S_{W_i} = \sum_{x_i \in \omega_1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{x_i \in \omega_2} (x_i - \mu_2)(x_i - \mu_2)^T + \ldots + \sum_{x_i \in \omega_c} (x_i - \mu_c)(x_i - \mu_c)^T \quad (4.4.6)$$

**Step 3: Constructing the LDA Lower Dimensional Space**

After calculating the between-class variance ($S_B$) and within-class variance ($S_W$), the LDA transformation matrix (W) can be computed as in Equation (4.4.7), which is called Fisher's criterion. This formula can be reformulated as in Equation (4.4.8).

$$argmax_W \frac{W^T S_B W}{W^T S_W W} \quad (4.4.7)$$

$$S_W \mathrm{W} = \lambda S_B W \quad (4.4.8)$$

where λ represents the eigenvalues of the transformation matrix (W). The solution of this problem can be obtained by calculating the eigenvalues ($\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_d\}$) and eigenvectors ($V = \{V_1, V_2, \ldots, V_d\}$) of $W = S_W^{-1} S_B$, if $S_W$ is non-singular (J. Lu et al., 2003; J. Ye et al., 2004).

The eigenvalues are scalar values, while the eigenvectors are non-zero vectors, which provide us with the information about the LDA space. The eigenvectors represent the directions of the new space, and the corresponding eigenvalues represent the magnitude of the eigenvectors (Zhu and Ogihara, 2006). Thus, each eigenvector represents one axis of the LDA space, and the associated eigenvalue represents the robustness of this eigenvector. Thus, the eigenvectors with the k highest eigenvalues are used to construct a lower dimensional space ($V_k$), while the other eigenvectors ($\{V_{k+1}, V_{k+2}, \ldots, V_d\}$) are neglected.

***Algorithm 4.3: Linear Discriminant Analysis (LDA)***

1) Given a set of n samples $\{x_i\}_{i=1}^{n}$, each of which is represented as a row of length d, and in this work LDA is applied on standardized data matrix $X_{n\times d}$ which is given by:

$$X = \begin{bmatrix} x_{(1,1)} & x_{(1,2)} & \cdots & x_{(1,d)} \\ \vdots & & \ddots & \vdots \\ x_{(n,1)} & x_{(n,2)} & \cdots & x_{(n,d)} \end{bmatrix} \qquad (4.4.10)$$

2) Compute the mean of each class $\mu_i$, $(1 \times d)$-dimensional as in Equation (4.4.2).
3) Compute the total mean of all data $\mu$, $(1 \times d)$- dimensional as in Equation (4.4.3).
4) Calculate between-class matrix $S_B$, $(d \times d)$- dimensional as follows:

$$S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T \qquad (4.4.11)$$

5) Compute within-class matrix $S_{W,}$ $(d \times d)$-dimensional, as follows:

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T \qquad (4.4.12)$$

where $x_{ij}$ represents the $i^{th}$ sample in the $j^{th}$ class.

6) From Equations (4.4.11 and 4.4.12), the matrix W that maximizing Fisher's formula is calculated as follows, $W = S_W^{-1}S_B$. The eigenvalues and eigenvectors of W are then calculated.
7) Sorting eigenvectors in descending order according to their corresponding eigenvalues. The first k eigenvectors are then used as a lower dimensional space ($V_k$).
8) Project all original samples (***X***) onto the lower dimensional space of LDA using the projection:

$$\boldsymbol{Y = XV_k}$$

# CHAPTER

# 5 RESULTS AND DISCUSSIONS

## 5.1 Principal Component Analysis (PCA)

Prior to PCA, the initial dataset of 10 parameters (variables) was checked using correlations inspection and suitability of the dataset for PCA model. Kaiser-Meyer-Olkin (KMO), Bartlett's tests and determinant measures statistics were used to assess suitability of green coffee beans dataset for basis of PCA. The adequacy evaluation was performed by Kaiser-Meyer-Olkin and Bartlett tests.

### 5.1.1 Results of One-Way ANOVA

Results of one-way ANOVA ($\alpha$ = 0.05) (Appendix A) shows significant for all except 5-pCoQA and 4,5-diCQA/3,4-diCQA variables, indicating the mean content of the sample coffees for each 8 variables differ significantly at regional level. On the other hand, Results of one-way ANOVA ($\alpha$ = 0.05) indicated that the mean content of sample coffees of each variable differ significantly at sub-regional level. Moreover, test of homogeneity of variances shows insignificant at p-value less than 0.05 for only 5-pCoQA at regional level (Appendix A).

### 5.1.2 Correlations Matrix Inspection

A correlation matrix was used as a basis for the application of the PCA analysis. We used Pearson`s linear correlation coefficients for the determination of the variables with highest impact on the components' extraction process. Tabachnick & Fidell (2001) said that if there are few correlations above 0.3, it is a waste of time carrying on with the analysis. However, clearly, we do not have that problem, and the correlation matrix showed good consistency of results. The result shows that a significant correlation was determined for all variables of the green coffee beans dataset with the exception of 5-pCoQA (Table 5.1.1). According to the correlation coefficients, variables with the strongest impact include 3-CQA, 5-CQa, 4-CQA, 5-FQA, 3,4-di-CQA, 3,5-diCQA, 4,5-diCQA, and 3,5-diCQA/4,5-diCQA. Strong significant positive correlation was established for all of the variables with the exception of 5-CQA, 3,5-diCQA, 3,5-diCQA/4,5-diCQA, and 4,5-diCQA/3,4-diCQA.

## Table 5.1.1: Correlation Matrix

| | | 3-CQA | 4-CQA | 3,4-diCQA | 4,5-diCQA | 5-FQA | 3,5-diCQA | 3,5-diCQA/4,5-diCQA | 5-CQA | 5-pCoQA | 4,5-diCQA/3,4-diCQA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3-CQA | Pearson Correlation | 1 | | | | | | | | | |
| | Sig. (2-tailed) | | | | | | | | | | |
| | N | 100 | | | | | | | | | |
| 4-CQA | Pearson Correlation | **.700**$^{**}$ | 1 | | | | | | | | |
| | Sig. (2-tailed) | .000 | | | | | | | | | |
| | N | 100 | 100 | | | | | | | | |
| 3,4-diCQA | Pearson Correlation | **.486**$^{**}$ | **.499**$^{**}$ | 1 | | | | | | | |
| | Sig. (2-tailed) | .000 | .000 | | | | | | | | |
| | N | 100 | 100 | 100 | | | | | | | |
| 4,5-diCQA | Pearson Correlation | **.446**$^{**}$ | **.400**$^{**}$ | **.664**$^{**}$ | 1 | | | | | | |
| | Sig. (2-tailed) | .000 | .000 | .000 | | | | | | | |
| | N | 100 | 100 | 100 | 100 | | | | | | |
| 5-FQA | Pearson Correlation | .221$^{*}$ | .347$^{**}$ | **.577**$^{**}$ | **.452**$^{**}$ | 1 | | | | | |
| | Sig. (2-tailed) | .027 | .000 | .000 | .000 | | | | | | |
| | N | 100 | 100 | 100 | 100 | 100 | | | | | |
| 3,5-diCQA | Pearson Correlation | -.524$^{**}$ | -.440$^{**}$ | -.473$^{**}$ | -.429$^{**}$ | -.205$^{*}$ | 1 | | | | |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .040 | | | | | |
| | N | 100 | 100 | 100 | 100 | 100 | 100 | | | | |
| 3,5-diCQA/4,5-diCQA | Pearson Correlation | -.560$^{**}$ | -.483$^{**}$ | -.589$^{**}$ | -.816$^{**}$ | -.387$^{**}$ | **.777**$^{**}$ | 1 | | | |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | | | | |
| | N | 100 | 100 | 100 | 100 | 100 | 100 | 100 | | | |
| 5-CQA | Pearson Correlation | -.146 | -.049 | **-.419**$^{**}$ | **-.520**$^{**}$ | -.272$^{**}$ | .213$^{*}$ | .401$^{**}$ | 1 | | |
| | Sig. (2-tailed) | .148 | .625 | .000 | .000 | .006 | .033 | .000 | | | |
| | N | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | | |
| 5-pCoQA | Pearson Correlation | -.047 | .024 | -.015 | -.023 | -.058 | .084 | .045 | .081 | 1 | |
| | Sig. (2-tailed) | .639 | .810 | .882 | .824 | .567 | .406 | .655 | .420 | | |
| | N | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| 4,5-diCQA/3,4-diCQA | Pearson Correlation | -.159 | -.192 | -.555$^{**}$ | .215$^{*}$ | -.195 | .147 | -.147 | .013 | -.064 | 1 |
| | Sig. (2-tailed) | .115 | .055 | .000 | .032 | .052 | .145 | .145 | .901 | .529 | |
| | N | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

### 5.1.3 Suitability of dataset for basis of PCA

Kaiser-Meyer-Olkin (KMO), Bartlett's and determinant measures statistics were used to assess suitability of green coffee beans dataset for basis of PCA. The adequacy evaluation was performed by Kaiser-Meyer-Olkin and Bartlett tests.

The Kaiser-Meyer-Olkin is the measure of sampling adequacy (MSA), which varies between 0 and 1. There studies (Hair et al., 2010; Pallant, 2007; Tabachnick and Fidell, 2007) suggested that if the Kaiser-Meyer-Olkin (KMO) is greater than 0.6 and the Bartlett's Test of Sphericity (BTS) must be significant at $\alpha < .05$, then factorability of the correlation matrix is assumed.

In other words, the KMO test and BTS determines whether the sampling was adequate to proceed with principal component analysis (Maat, Zakaria, Nordin, & Meerah, 2011). In addition, a few steps need to be taken into the account by the researcher was the anti-image correlation for all variables must the acceptable level, above 0.5 (Coakes, Steed, Coakes, & Steed, 2003; Hair et al., 2010). Besides, the results provided for all items had a communality that was above 0.3 (Tabachnick & Fidell, 2007).

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's test of Sphericity were conducted on the green coffee beans dataset. The ten variables of our dataset were subjected to principal component analysis (PCA) using SPSS software. Prior to constructing PCA model, the suitability of dataset was assessed. The sampling is adequate if the value of KMO test value is greater than 0.5 (Field, 2000; Kaiser, 1974). Accordingly, the Kaiser Meyer-Olkin value is 0.618 (Table 5.1.2), exceeding the recommended minimum value of 0.5 (Kaiser 1970, 1974). It was also supported by Bartlett's test of Sphericity and found the results are significant with Chi-Square value of 761.530. Moreover, the KMO value for each variable (with the exception of 5-yCQA and 4,5-diCQA/3,4-diCQA) was above 0.5 (Table 5.1.3).

Therefore, by removing the two parameters (variables), 5-yCQA and 4,5-diCQA/3,4-diCQA, the results indicated that the obtained data were suitable for performing principal component analysis and that the suggested model is adequate.

**Table 5.1.2: Kaiser-Meyer-Olkin (KMO) and Bartlett's Test of Sphericity**

**KMO and Bartlett's Test[a]**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .618 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 761.530 |
| | df | 45 |
| | Sig. | .000 |

a. Based on correlations

**Table 5.1.3: Kaiser-Meyer-Oklin (KMO) value for each variable**

| Variable | KMO value |
|---|---|
| 3-CQA | .788[a] |
| 5-CQA | .829[a] |
| 4-CQA | .783[a] |
| 5-pCoQA | .100[a] |
| 5-FQA | .835[a] |
| 3,4-diCQA | .550[a] |
| 3,5-diCQA | .653[a] |
| 4,5-diCQA | .581[a] |
| 3,5-diCQA/4,5-diCQA | .724[a] |
| 4,5-diCQA/3,4-diCQA | .205[a] |

a. Measures of Sampling Adequacy(MSA)

Given these overall indicators, 5-pCoQA and 4,5-diCQA/3,4-diCQA variables were removed and principal component analysis was then conducted with 8 variables (i.e. 7 compounds and 3,5-diCQA/4,5-diCQA ratio) of Green Coffee beans dataset with Varimax rotation. The minimum factor loading cut off point of this study was 0.4.

Accordingly, the updates overall Kaiser Meyer-Olkin value was 0.701 and KMO for each variable also is above 0.6, exceeding the recommended minimum value of 0.5 (Kaiser 1970,

1974). In addition, the strength of the relationships can be measured by a Bartlett Test of Sphericity. The Bartlett's Test of Sphericity was statistical significance (p-value of .000) with Chi-Square value of 496.606 and degree of freedom, df= 28.

The value of determinant matrix is 0.041, exceeding the recommended minimum value of 0.000001, and hence, the results indicated that the obtained data were suitable for performing principal component analysis and that the suggested model is adequate.

### 5.1.4 Component Extraction

Component (Factor) extraction is the way of defining a small number of components that can be used to best signify the whole of the relationship between variables. There are a number of methods that might be used to help in making decision regarding determining the smaller number of factors that should be retained. The following are the well-known techniques for factor extraction.

(1) Kaiser's criterion, (2) Scree test (Catell, 1996), (3) Parallel Analysis (Horn, 1965)

One of the most frequently used methods is known as the Eigen value rules or the Kaiser's criteria. Under this criteria, components with an Eigen value larger than 1 are retained, or factors which explain a total of 70-80% of the variance is retained, or do a screen test (Catell, 1996), plotting a plot and review the plot based on Catell criteria; as according to him keep all factors above the elbow. Similarly, Horn (1965) Horn's parallel analysis is also the best way to factor extraction. Under this method of factor extraction the Eigen values are compared with the values obtained from another statistical program called Monte Carlo PCA for parallel analysis and those factors are retained whose actual Eigen value from principal component analysis (PCA) more than that of the value are obtained from Monte Carlo PCA.

For the purpose of this study, this thesis used the Eigen value rules (i.e. the Kaiser's criteria) to determine the number of components (known as PCs) for the principal component analysis (PCA). Accordingly, it was found that three components had values of eigenvectors greater than 1, which determined the choice of three principal components for our analysis. Thus, based on the Eigenvalue rule (or Kaiser Criterion), three components with eigenvalue of greater than 1 retained for further analysis. The total variance of 89.3% is achieved from these three principal components. The first component's eigenvalue is equal to 8.71 and explains 55.7% of the variance in the original dataset. The second component's eigenvalue is

equal to 4.12 and explains 26.3% of the variance, and the third component's eigenvalue is equal to 1.138, which explains 7.3% of the variance as shown in Table 8.

Given these overall indicators, principal component analysis was then conducted with 8 variables with Varimax rotation. The minimum component loading cut off point for this study was 0.4.

**Table 5.1.4: Eigenvalues and eigenvectors of the covariance matrix and total variance explained**

| | Principal Component | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 3-CQA | 0.131 | 0.302 | 0.373 | -0.524 | -0.211 | 0.627 | -0.093 | 0.031 |
| 5-CQA | -0.315 | 0.180 | 0.024 | 0.061 | -0.020 | 0.012 | 0.014 | 0.004 |
| 4-CQA | 0.097 | 0.301 | 0.510 | -0.370 | 0.206 | -0.719 | -0.060 | 0.019 |
| 5-FQA | 0.133 | 0.087 | 0.433 | 0.559 | 0.822 | 0.346 | -0.421 | 0.124 |
| 3,4-diCQA | 0.206 | 0.172 | 0.323 | 0.218 | 0.292 | 0.181 | 1.272 | -0.583 |
| 3,5-diCQA | -0.175 | -0.378 | 0.344 | 0.041 | -0.143 | 0.000 | -0.045 | -0.139 |
| 4,5-diCQA | 0.240 | 0.143 | 0.305 | 0.471 | -0.492 | -0.077 | 0.206 | 0.653 |
| 3,5-diCQA/4,5-diCQA | -0.233 | -0.301 | -0.017 | -0.309 | 0.329 | 0.073 | 0.299 | 0.545 |
| Eigenvalue | 8.706 | 4.116 | 1.138 | 0.696 | 0.477 | 0.259 | 0.175 | 0.067 |
| Variance (%) | 55.7 | 26.3 | 7.3 | 4.5 | 3.1 | 1.7 | 1.1 | 0.4 |
| Cumulative (%) | 55.7 | 82.0 | 89.3 | 93.7 | 96.8 | 98.5 | 99.6 | 100.0 |

### 5.1.5 The PCA model

The communality is the proportion of common variance within a variable. Therefore, before extraction, all of the variance associated with a variable assumed to be common variance. PCA work on the assumption that all the variance associated with a variables supposed to be 1 before factors extraction. Thus, this communality table/matrix gives information about how much of the variance in each item is explained. Low value 0.3 indicates that the item does not fit well with another item in its component. It also means that before extraction there are many factors, therefore all the communalities are 1 and all of the variance is explained by all factors. After factor extraction some of the factors are thrown away as a result some information is lost. Thus, the retained factors after factor rotation cannot explain all of the variance presents in the data, nevertheless they can explain some. Therefore, the communalities represent the degree of variance in each variable that can be explained by the retained factors after extraction. Accordingly, the communalities of this study were

determined for each item. The communalities of the variables were range from 0.615 to 0.874, table below.

A table of loadings should be examined next, as it shows which variables have high loadings (positive or negative) on each principal component, that is, which variables contribute most strongly to each PC. Examining this table can give a good sense of what each principal component represents, in terms of the original data. A positive loading means that a variable correlates positively with the principal component; a negative loading indicates a negative correlation.

Component matrix displays the un-rotated loading of each of the items on components. Before making decision to retain two components or three components you have to look for this matrix. This shows the item loading on each component. It is better considering more number of items load above 0.4 on first and second components, while lesser number of items load on the third and forth components. Accordingly, the result from tale below shows seven items, having component loadings greater than 0.5, load on component 1, three loads on component 2, and one loads on component 3. This recommends that choosing three components being more suitable for further investigations.

**Table 5.1.5: Component and communality matrix**

| Variable | Component | | | Communalities | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | Initial | Extraction |
| 3,4-diCQA | **.862** | -.155 | .206 | 1.000 | .817 |
| 3,5-diCQA/4,5-diCQA | **-.854** | -.041 | .328 | 1.000 | .745 |
| 4,5-diCQA | **.842** | -.237 | -.221 | 1.000 | .796 |
| 3-CQA | **.702** | **.565** | -.072 | 1.000 | .874 |
| 4-CQA | **.675** | **.540** | .223 | 1.000 | .809 |
| 3,5-diCQA | **-.663** | -.278 | .313 | 1.000 | .615 |
| 5-CQA | **-.500** | **.595** | .374 | 1.000 | .813 |
| 5-FQA | **.616** | -.328 | **.622** | 1.000 | .839 |

Extraction Method: Principal Component Analysis.

### 5.1.6 Component Rotation and Interpretation

It can be hard to name and interpret the components after extraction based on their factor loadings because PCA criteria is that the first component account for the maximum part of the variance, so that understanding of the components might be hard. Therefore, to interpret them, the rotation of components assists in this process. Accordingly, component rotation changes the pattern of the components (as shown in component matrix) and increases the understanding of each component, by presenting the pattern of loadings in a manner that is easier to interpret and understand (Pallant, 2013: 184).

There are two types of rotation: *Orthogonal Rotation* (uncorrelated) and *Oblique Rotation* (correlated). In this study orthogonal rotation was used. In orthogonal rotation there is no correlation between the extracted factors. Here, the results are *rotated component matrix* and *component transformation matrix*. The former matrix represents the *pattern of loadings* of the original indicators on retained or extracted components, while the second matrix provides information regarding the *angle of rotation*. Rotation matrix comprises similar information as the component matrix with the exception that, it is calculated after factor rotation. The format of the matrix should be considered, the suppression of loadings greater than 0.4 and the ordering variable by loading size.

To easily understand and in support of interpretation of these three components, Varimax rotation method was conducted. Percentage variance in Extraction Sums of Squared Loading and Rotation Sums of Squared Loadings are the same, which explains 89.293%. Varimax rotation has resulted the percentage of variance for Component 1 has been changed from 55.683 to 15.767%, percentage of variance for Component 2 has been changed from 26.329% to 25.269%, while percentage of variance for Component 3 has been changed from 7.281% to 48.257%.

**Table 5.1.6: Total variance explained**

| Component | Initial Eigenvalues[a] | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 8.706 | 55.683 | 55.683 | 8.706 | 55.683 | 55.683 | 2.465 | 15.767 | 15.767 |
| 2 | 4.116 | 26.329 | 82.013 | 4.116 | 26.329 | 82.013 | 3.951 | 25.269 | 41.037 |
| 3 | 1.138 | 7.281 | 89.293 | 1.138 | 7.281 | 89.293 | 7.545 | 48.257 | 89.293 |
| 4 | .696 | 4.454 | 93.747 | | | | | | |
| 5 | .477 | 3.049 | 96.796 | | | | | | |
| 6 | .259 | 1.658 | 98.454 | | | | | | |
| 7 | .175 | 1.118 | 99.572 | | | | | | |
| 8 | .067 | .428 | 100.000 | | | | | | |

Extraction Method: Principal Component Analysis.

The Rotated Component Matrix (Table 5.1.7) represents the rotated component loadings, which are the correlations between the variables and the components (PCs). The component column represents the rotated components that have been extracted out of the total component. These are the core components, which have been used as the final component after data reduction. According to the grouping of the components, each group of components is named which will represent the grouped components.

The above matrix gives the correlation of the variables with each of the extracted components. Usually, each of the variables is highly loaded in one component and less loaded towards the other components. To identify the variables, included in each component, the variable with the value maximum in each row is selected to be part of the respective component. The values have been high lightened in each of the rows to group the 8 variables into 3 core components.

It is visible from the results that the three main components covered 89.3% of the total variance and confirmed for the graphic interpretation of the results. Hence, the first three PCs are the main components sufficient for the application of the PCA model.

The variables 4-CQA, 3-CQA, 3,4-diCQA, 4,5-diCQA, and 5-CQA had a positive component loading and 3,5-diCQA, 3,5-diCQA/4,5-diCQA, and 5-FQA had a negative component loading in the formation of PC1. The variables: 3,5-diCQA, 3,5-diCQA/4,5-

diCQA, 5-CQA, and 4-CQA had a positive component loading and 3-CQA, 3,4-diCQA, 5-FQA, and 4,5-diCQA had a negative component loadings in the formation of both PC2. Similarly, 3-CQA, 4-CQA, 3,4-diCQA, 5-FQA, 3,5-diCQA, and 4,5-diCQA had a positive component loadings while 3,5-diCQA/4,5-diCQA and 5-CQA had a negative component loadings in the formation of both PC3 (Table 5.1.7).

Accordingly, most of the variables connected with each factor were well defined after rotation with Varimax. The result revealed that the three variables: 3-CQA, 4-CQA, and 3,5-diCQA with the highest component loadings on the first component and lowest loadings on others were associated with the first component (PC 1). Similarly, three variables: 4,5-diCQA, 3,5-diCQA/4,5-diCQA, and 5-CQA with highest loadings on PC 2 and lowest loadings on others were associated with the second component (PC 2), while two variables 5-FQA and 3,4-diCQA with the highest loadings on PC 3 and lowest loadings on other was associated with the third component (PC 3) (Table 5.1.7).

Hence, the results from our study have demonstrated the reliability of the PCA method in the process of determination of the main components influencing the identification of the most discriminating variables (compounds) for the green coffee beans.

**Table 5.1.7: Rotated component matrix**

**Rotated Component Matrix[a]**

| | Raw | | | Rescaled | | |
|---|---|---|---|---|---|---|
| | Component | | | Component | | |
| | 1 | 2 | 3 | 1 | 2 | 3 |
| 3-CQA | .201 | -.023 | .027 | **.890** | -.101 | .118 |
| 4-CQA | .172 | .015 | .073 | **.819** | .073 | .348 |
| 3,5-diCQA | -.121 | .069 | .000 | **-.680** | .390 | .001 |
| 5-CQA | .015 | .179 | -.031 | .069 | **.847** | -.147 |
| 4,5-diCQA | .093 | -.170 | .076 | .403 | **-.737** | .328 |
| 3,5-diCQA/4,5-diCQA | -.126 | .128 | -.032 | -.631 | **.644** | -.160 |
| 5-FQA | .020 | -.038 | .207 | .087 | -.167 | **.916** |
| 3,4-diCQA | .112 | -.117 | .171 | .427 | -.448 | **.653** |

**Rotated Component Matrix[a]**

| | Raw | | | Rescaled | | |
|---|---|---|---|---|---|---|
| | Component | | | Component | | |
| | 1 | 2 | 3 | 1 | 2 | 3 |
| 3-CQA | .201 | -.023 | .027 | **.890** | -.101 | .118 |
| 4-CQA | .172 | .015 | .073 | **.819** | .073 | .348 |
| 3,5-diCQA | -.121 | .069 | .000 | **-.680** | .390 | .001 |
| 5-CQA | .015 | .179 | -.031 | .069 | **.847** | -.147 |
| 4,5-diCQA | .093 | -.170 | .076 | .403 | **-.737** | .328 |
| 3,5-diCQA/4,5-diCQA | -.126 | .128 | -.032 | -.631 | **.644** | -.160 |
| 5-FQA | .020 | -.038 | .207 | .087 | -.167 | **.916** |
| 3,4-diCQA | .112 | -.117 | .171 | .427 | -.448 | **.653** |

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

From the loading plot below, five variables 4-CQA (qa4), 3-CQA (qa3), 3,4-diCQA (qa34), 4,5-diCQA (qa45), and 5-CQA (qa5) for located on the positive side the first component (PC1). On the other hand, three variables 3,5-diCQA (qa35), 3,5-diCQA/4,5-diCQA (qa35to45), and 5-FQA (qa5f) are positioned on the positive side of PC2 and on the negative side PC1.

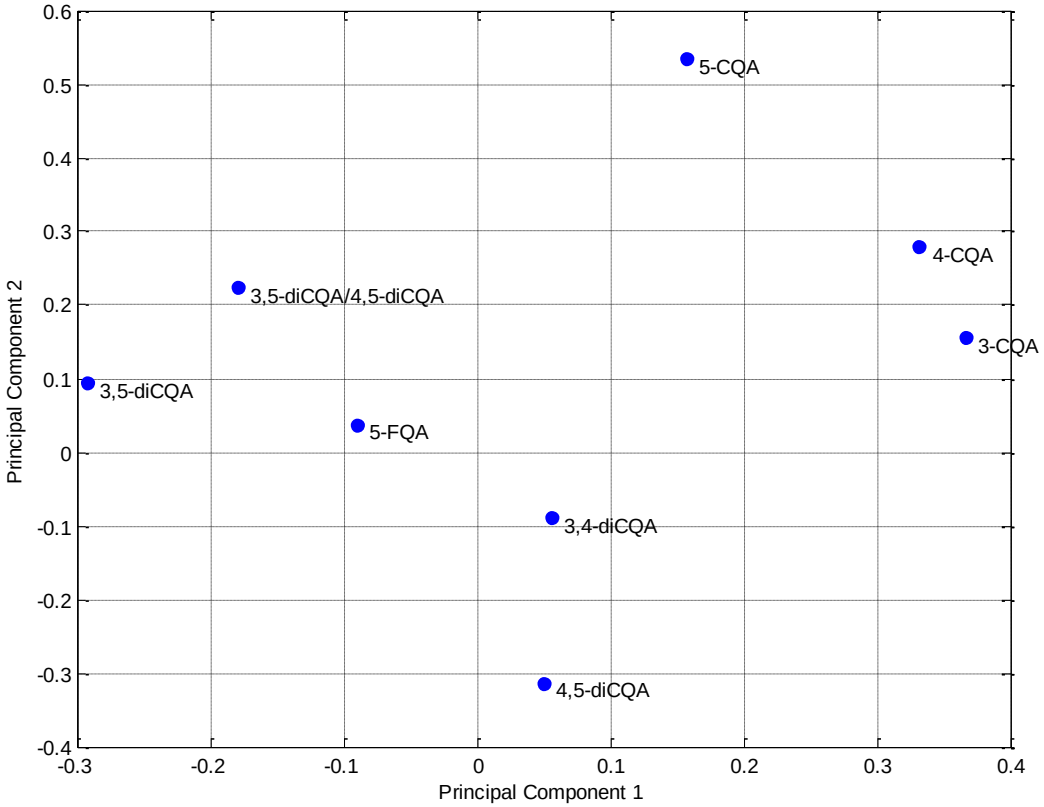



**Figure 5.1.1: Loading Plot of the first two PCs in Rotated Space**

### 5.1.7 PC Score Interpretation

Score plot of the first two PCs gives information about grouping of samples. Samples in the score plot are color coded, where the green represents the sample from east geographical region, red represents the samples from the west, blue represents the samples from northwest, and black represents the samples from the south geographical region.

The score plot gives information about grouping, outliers and other patterns in the sample dataset. Analyzing the score plot given in Figure 5.1.2, it is clear that this data matrix can be grouped (identifies) most of the green coffee samples into their geographical origin. The green coffee samples from the east are towards the right-bottom part of the plot and the samples from the northwest are towards the right-top part of the plot. On the other hand, most of the green coffee samples from the west are towards the left-top part of the plot and some of the samples from the south are towards the left-bottom part of the plot while some samples at the middle and top part.

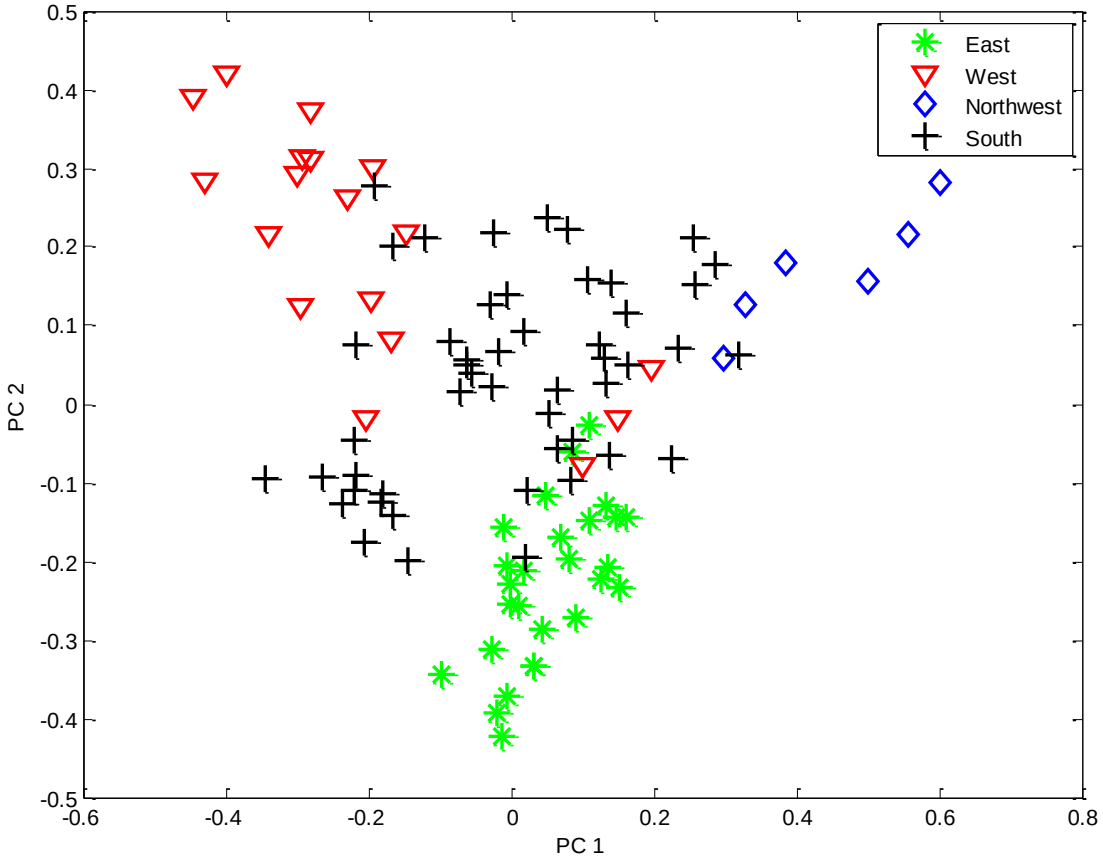


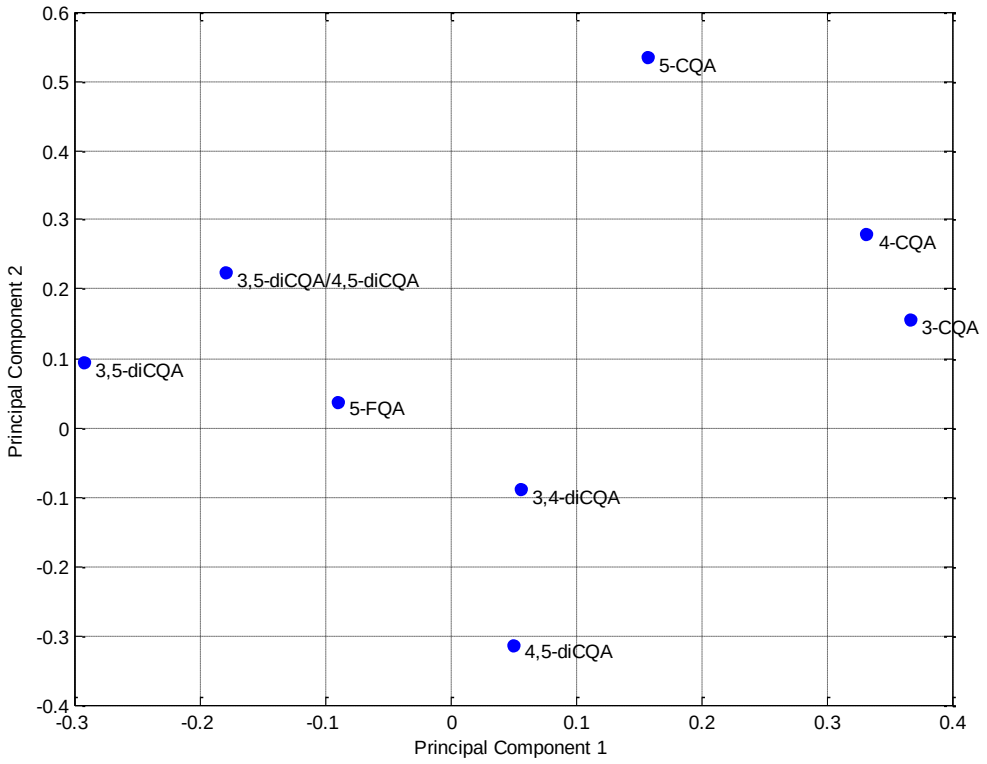

**Figure 5.1.2: Score plot of the first two PCs**

From the superposition of the loadings plot and score plot, 3-CQA was located in the same area with the Northwest coffees. These coffee samples are located in the positive direction both of the first and second component, with higher influences of component one. Hence, northwest coffees are characterized by mainly 3-CQA. It is also confirmed from the bi-plot (Figure 5.1.4) and hence, 3-CQA distinguished Northwest coffees from the other regional green coffee beans.

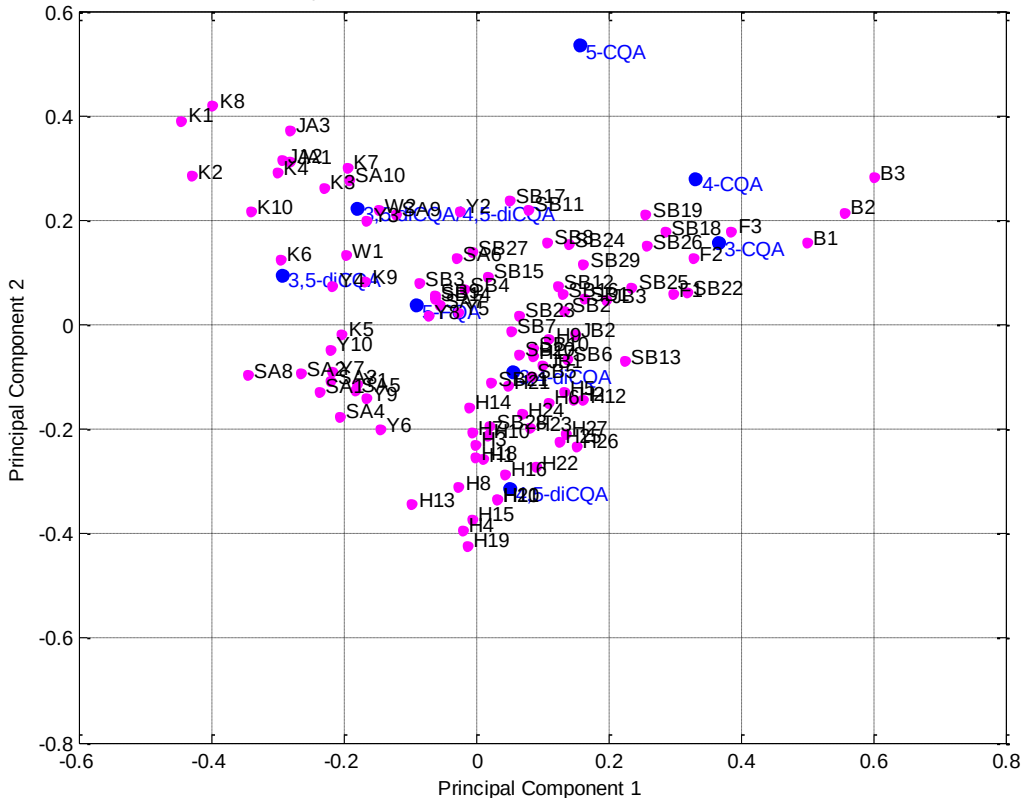
Similarly, from the loadings plot, score plot and bi-plot, the green coffee samples from the East geographical region were distributed in the direction of 4,5-diCQA and 3,4-diCQA compounds. Hence, coffees from east region are characterized and distinguished from other regional coffees by the larger influence of 4,5-diCQA and lesser 3,4-diCQA compounds.

The loadings plot and scores plot of PCA indicated that green coffee beans from West, with the exception of Jimma B, can be characterized by their greater contents of 3,5-diCQA/4,5-diCQA concentration ratio. On the other hand, some coffee samples from South were

distributed in the same direction with 3,5-diCQA/4,5-diCQA concentration ratio, and some located in the same direction with the 5-FQA compound.




**(a)**




**(b)**

**Figure 5.1.3: Superposition of the loadings plot (a) and score plot (b) of PC1 and PC2**

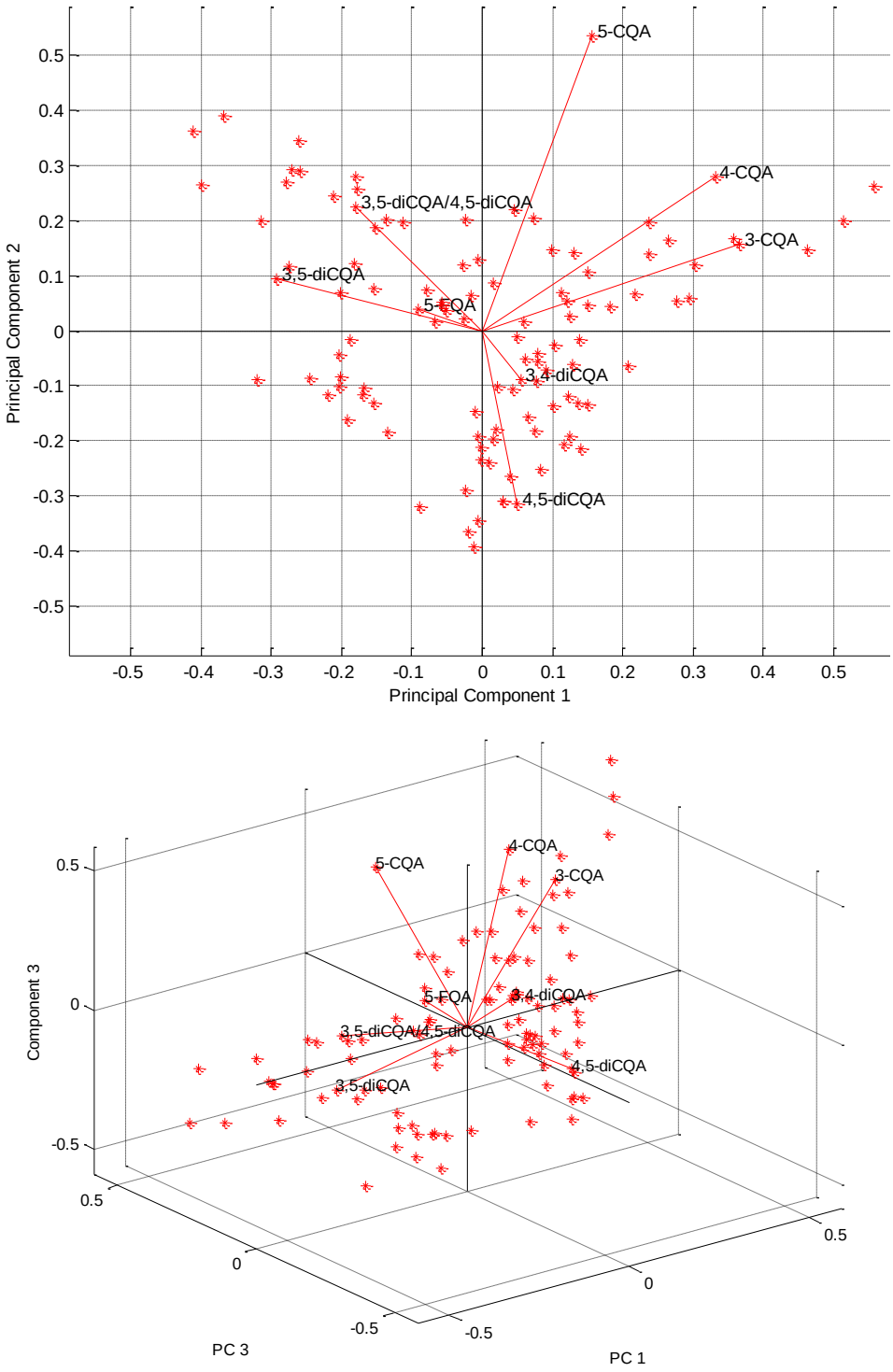


**Figure 5.1.4: The Bi-plot of loadings and scores for the first 2 PCs of green coffee beans**

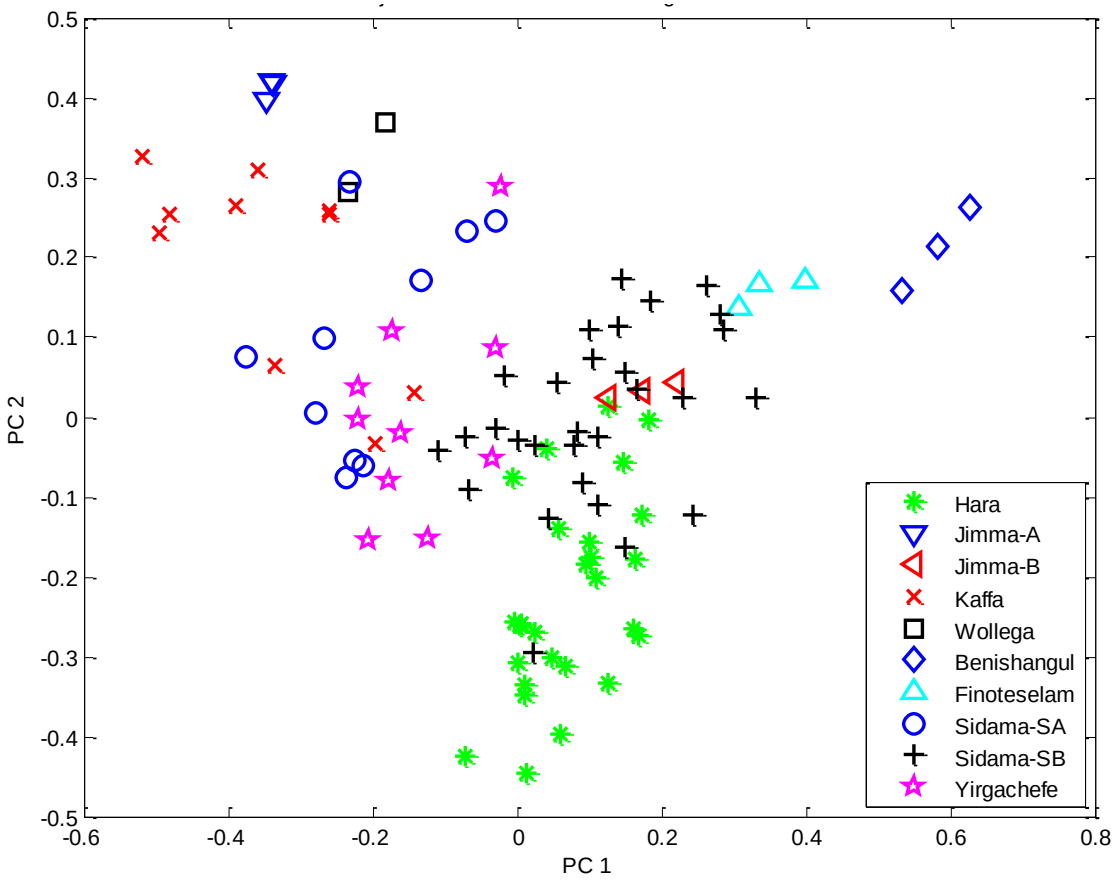



**Figure 5.1.6: PCA score on the first two PCs at sub-regional level**

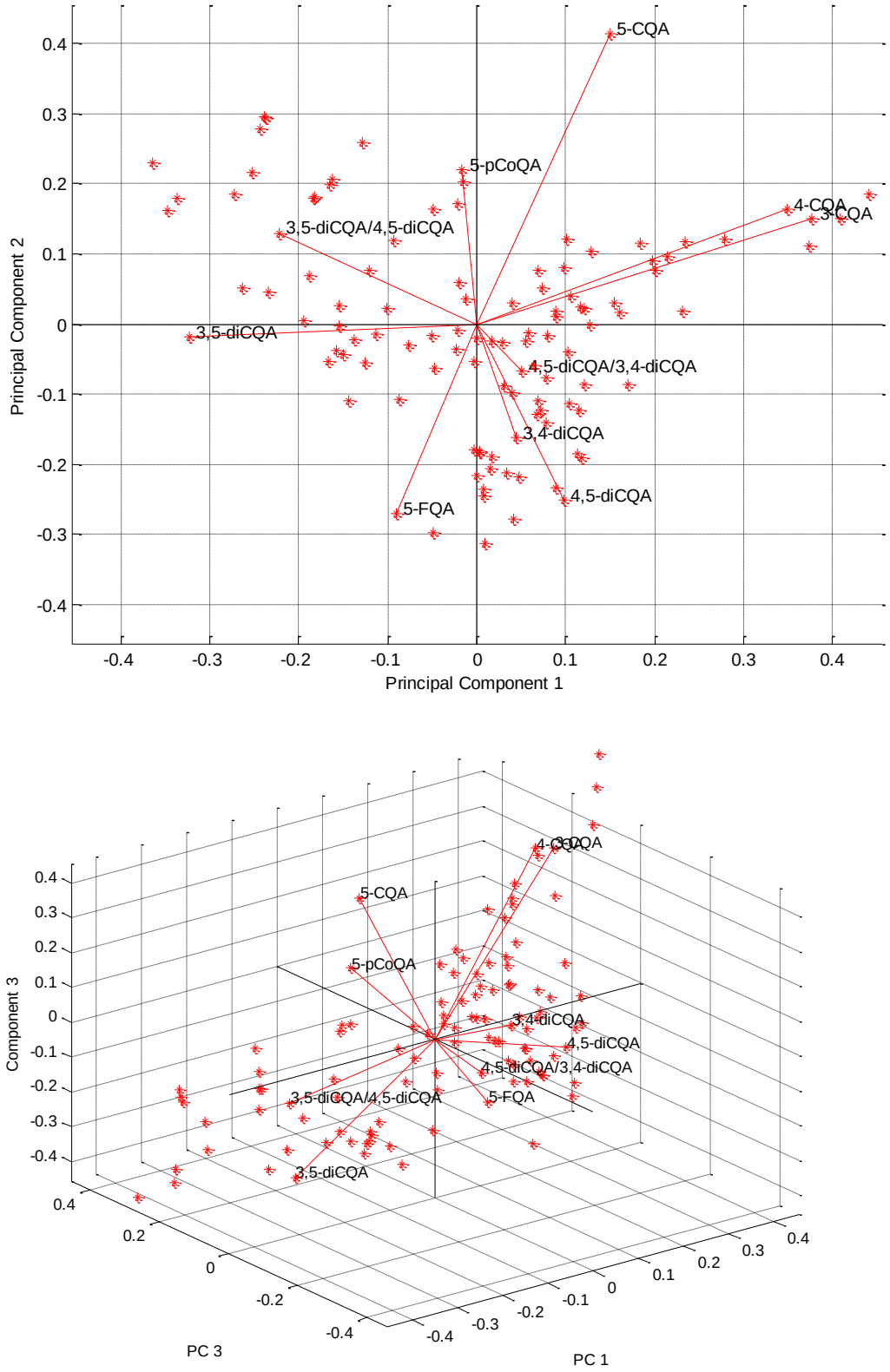



**Figure 5.1.7: The bi-plots PCA loadings and scores on the first two PCs at sub-regional level**

## 5.2 Linear Discriminant Analysis (LDA)

### 5.2.1 Classification at Regional Level

Discriminant analysis is used to predict a group membership, so an examination of whether there are any significant differences between groups on each of the variables using group means and ANOVA results need to be done first. The Group Statistics and Tests of Equality of Group Means tables provide this information. An examination of the group means and standard deviations can also be helpful in obtaining a rough idea of variables that may be important. Accordingly, the variations in the mean values of the variables between the groups suggest these may be good discriminators (Appendix B).

Table 5.2.1 provides statistical evidence of significant differences between means of East, West, Northwest and South Region categories for all variables with 3-CQA, 5-CQA, 4-CQA, 5-FQA, 3,4-diCQA, 3,5-diCQA, 4,5-diCQA, and 3,5-diCQA/4,5-diCQA producing very high value F's. The smaller the Wilks' lambda shows the more important the variable to the discriminant function. For this study, Wilks' lambda is significant by the F test for all variables with the exception of 5-pCoQA and 4,5-diCQA/3,4-diCQA.

**Table 5.2.1: Tests of equality of group means table**

| | Wilks' Lambda | F | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| 3-CQA | .509 | 30.825 | 3 | 96 | .000 |
| 5-CQA | .618 | 19.793 | 3 | 96 | .000 |
| 4-CQA | .687 | 14.581 | 3 | 96 | .000 |
| 5-pCoQA | .992 | .255 | 3 | 96 | .857 |
| 5-FQA | .710 | 13.040 | 3 | 96 | .000 |
| 3,4-diCQA | .504 | 31.462 | 3 | 96 | .000 |
| 3,5-diCQA | .712 | 12.948 | 3 | 96 | .000 |
| 4,5-diCQA | .317 | 68.886 | 3 | 96 | .000 |
| 3,5-diCQA/4,5-diCQA | .463 | 37.079 | 3 | 96 | .000 |
| 4,5-diCQA/3,4-diCQA | .974 | .867 | 3 | 96 | .461 |

#### 5.2.1.1 Summary of Canonical Discriminant Functions at Regional levels

This gives information on each of the discriminate functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus 1. In this case, four groups were used, namely East, West, Northwest and South, so only three functions are displayed. An eigenvalue provides information on the proportion of variance explained. A large eigenvalue is associated with a strong function. The canonical correlation measures the degree of association between the discriminant function and the predictors. In Table 5.2.2, a canonical correlation of 0.895 suggests that the model explains 80.10% (i.e.$0.895^2$ *100) of the variation in the grouping variable (Burns and Burns, 2008).

**Table 5.2.2: Eigenvalues table (region)**

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|---|---|---|---|---|
| 1 | 4.011[a] | 69.1 | 69.1 | .895 |
| 2 | 1.094[a] | 18.8 | 87.9 | .723 |
| 3 | .701[a] | 12.1 | 100.0 | .642 |

a. First 3 canonical discriminant functions were used in the analysis.

Wilks' lambda is of great analytic importance. It is used to measure how well each function separates cases into groups. The smaller values of Wilks' lambda indicate greater discriminatory ability of the function. The significance of the discriminant function is indicated by Wilks' lambda and provides the proportion of total variability not accounted. The result of this study shows a highly significant function (sig. =0.000) and 5.6% unexplained (i.e. Wilks' lambda value of 0.056), which means using that combination of weights on the variables of these study leaves about 5.6% of the variance in the four regions (groups) unexplained.

Moreover, the Chi-Square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value indicates that the discriminant function does better than chance at separating the groups.

**Table 5.2.3: Wilks' lambda table (region)**

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 through 3 | .056 | 270.943 | 18 | .000 |
| 2 through 3 | .281 | 119.449 | 10 | .000 |
| 3 | .588 | 49.957 | 4 | .000 |

Table 5.2.4 provides an index of the importance of each predictor with the sign indicating the direction of the relationship. Coefficients with large absolute values correspond to variables with greater discriminating ability. The result shows the variables 4,5-diCQA compound, 3,5-diCQA/4,5-diCQA ratio, and 3-CQA compound were the strongest predictors for the first three functions, followed by 3,5-diCQA compounds. These three variables with large coefficients stand out as those that strongly predict the green coffee beans to geographical categories (groups).

The other two compounds (variables) 5-CQA and 5-FQA scored less and have lower contribution in predicting the model, whereas the remaining variables 4-CQA, 3,4-diCQA, 5-pCoQA and 4,5-diCQA/3,4-diCQA ratio were excluded from the LDA model constructing at the regional level. The forward stepwise variable selection method was used to select the most relevant variables (compounds) for the discrimination of regional coffees.

**Table 5.2.4: Standardized canonical discriminant function coefficients (region)**

| | Function | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 3-CQA | -.507 | -.822 | .527 |
| 5-CQA | -.411 | .036 | .211 |
| 5-FQA | .276 | .082 | -.543 |
| 3,5-diCQA | -.944 | -.284 | -.957 |
| 4,5-diCQA | 1.705 | .482 | 1.127 |
| 3,5-diCQA/4,5-diCQA | 1.407 | .926 | 2.025 |

The structure matrix correlations are considered more accurate than the Standardized Canonical Discriminant Function Coefficients; it also indicates the relative importance of the predictors. The structure matrix table (Table 5.2.5) shows the correlations of each variable with each of the discriminate function. These Pearson coefficients are structure coefficients (or discriminant loadings), which have the same function like component loadings in PCA. The largest absolute coefficients for each discriminate function determine how each function is to be named. Similarly, as in component loadings, coefficients with absolute value ≥ 0.30 are used as the cut-off between important and less important variables. The result shows that 4,5-diCQA compound is the most important predicator of the first function followed by 5-CQA. Similarly, the compound 3-CQA is the most important predicator of the second function followed by 3,5-diCQA/4,5-diCQA concentration ratio and 3,5-diCQA compound, while 5-FQA compound is the highest predicator of the third function of the LDA model.

**Table 5.2.5: Structure matrix table (for Region)**

| | Function | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| 4,5-diCQA | **.682*** | -.511 | -.042 |
| 5-CQA | **-.379*** | .131 | .185 |
| 5-pCoQA[a] | -.272* | .057 | .004 |
| 4,5-diCQA/3,4-diCQA[a] | .263* | .008 | -.048 |
| 3-CQA | .061 | **-.897*** | .310 |
| 3,5-diCQA/4,5-diCQA | -.355 | **.720*** | .351 |
| 3,5-diCQA | -.147 | **.535*** | .084 |
| 4-CQA[a] | .180 | -.452* | .073 |
| 3,4-diCQA[a] | .204 | -.352* | -.068 |
| 5-FQA | .263 | -.136 | **-.396*** |

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

From the canonical discriminant function coefficients table, we have unstandardized coefficients operate like unstandardized b (in Multiple Regression) coefficients and are used to create the actual prediction equation which are used to classify new cases. So, from the unstandardized coefficients (Table 5.2.6), the prediction equation for discriminant functions should be like this:

D1 = -4.518 - 0.688*3-CQA - 0.191*5-CQA + 0.397*5-FQA - 0.637*3,5-diCQA + 3.138*4,5-diCQA + 1.754*3,5-diCQA/4,5-diCQA

D2 = -1.362 – 1.117*3-CQA + 0.017*5-CQA + 0.118*5-FQA - 0.192*3,5-diCQA + 0.88874,5-diCQA + 1.154*3,5-diCQA/4,5-diCQA

D3 = -10.266 + 0.716*3-CQA + 0.098*5-CQA - 0.780*5-FQA - 0.646*3,5-diCQA + 2.074,5-diCQA + 2.525*3,5-diCQA/4,5-diCQA

The discriminant function coefficients *b* or standardized form *beta* both show the partial contribution of each variable to the discriminate function controlling for all other variables in the equation. They can be used to assess each predicator variables unique contribution to the discriminate function and thus, provide information on the relative importance of each variable. Hence, substituting the values for a specific case will compute the canonical variable score for the function. When there are more than two groups, the number of canonical variables is *k-1* (where *k* is the number of groups) or *p* (the number of variables), whichever is smaller (Burns and Burns, 2008).

**Table 5.2.6: Canonical Discriminant Function Coefficients table (for Region)**

| | **Function** | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| 3-CQA | -.688 | -1.117 | .716 |
| 5-CQA | -.191 | .017 | .098 |
| 5-FQA | .397 | .118 | -.780 |
| 3,5-diCQA | -.637 | -.192 | -.646 |
| 4,5-diCQA | 3.138 | .887 | 2.074 |
| 3,5-diCQA/4,5-diCQA | 1.754 | 1.154 | 2.525 |
| (Constant) | -4.518 | -1.362 | -10.266 |

Unstandardized coefficients

#### 5.2.1.2 Regional Classification Statistics

Summary of number and percent of groups classified correctly and incorrectly are presented in classification table below. The result shows that all 27 cases from the East group and 6 cases from the Northwest group were 100% correctly classified. Among the 49 cases of the South group, 48 (98.0%) of them were correctly classified, while only one (2.0%) of them was incorrectly classified into the West group. Similarly, of the 18 samples from the West group, 12 (66.7%) were correctly classified and 6 (33.3%) were incorrectly classified into the South group (Table 5.2.7).

The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications. This produces a more reliable function and the idea behind it is that one should not use the case to be predicted as part of the categorization process. From the cross-validated table, all of the cases from the East and Northwest were 100% correctly classified. On the other hand, 48 (98.0%) samples from South and 12 (66.7%) samples of West were correctly classified, while one (2.0%) samples of South and 6 (33.3%) samples of West were incorrectly classified into the West and South, respectively.

In general, the classification results revealed that 93.0% of the green coffee beans cases were classified correctly into ‘East’ or ‘West’ or ‘Northwest’ or ‘South’ geographical areas of the country. Regarding cross-validation, the study applied the leave-one-out cross-validation, in which each case is classified by the functions derived from all cases other than that case. The result shows that 93.0% of cross-validated grouped cases correctly classified

**Table 5.2.7: Regional Classification results**

**Classification Results[b,c]**

| | | Region Category | Predicted Group Membership | | | | Total |
|---|---|---|---|---|---|---|---|
| | | | East | West | North-West | South | |
| Original | Count | East | 27 | 0 | 0 | 0 | 27 |
| | | West | 0 | 12 | 0 | 6 | 18 |
| | | North-West | 0 | 0 | 6 | 0 | 6 |
| | | South | 0 | 1 | 0 | 48 | 49 |
| | % | East | 100.0 | .0 | .0 | .0 | 100.0 |
| | | West | .0 | 66.7 | .0 | 33.3 | 100.0 |
| | | North-West | .0 | .0 | 100.0 | .0 | 100.0 |
| | | South | .0 | 2.0 | .0 | 98.0 | 100.0 |
| Cross-validated[a] | Count | East | 27 | 0 | 0 | 0 | 27 |
| | | West | 0 | 12 | 0 | 6 | 18 |
| | | North-West | 0 | 0 | 6 | 0 | 6 |
| | | South | 0 | 1 | 0 | 48 | 49 |
| | % | East | 100.0 | .0 | .0 | .0 | 100.0 |
| | | West | .0 | 66.7 | .0 | 33.3 | 100.0 |
| | | North-West | .0 | .0 | 100.0 | .0 | 100.0 |
| | | South | .0 | 2.0 | .0 | 98.0 | 100.0 |

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 93.0% of original grouped cases correctly classified.

c. 93.0% of cross-validated grouped cases correctly classified.

As it can be seen from the scatter plot (see below Figure 5.2.1), the first discriminant function clearly separate samples of East (Harar) coffees from the other region samples, whereas the second discriminant function clearly separates samples of Northwest coffees from the other region samples.

Moreover, the coefficients of canonical discriminant function (Table 5.2.6) indicates the first function was highly influenced to the positive side by 4,5-diCQA (3.138). Hence, East

coffees are separated from other regional coffees by their higher contents of 4,5-diCQA compound.

Similarly, Table 5.2.6 confirmed that the second function was highly influenced to the negative side by 3-CQA (-1.117). Hence, Northwest coffees are separated from other regional coffees by their higher content of 3-CQA compound.

On the other hand, using scatter plots of the three functions each other, most coffees from the West separated by 3,5-diCQA/4,5-diCQA concentration ratios, and some of the coffees from South separated by 3,5-diCQA and some by 5-FQA.

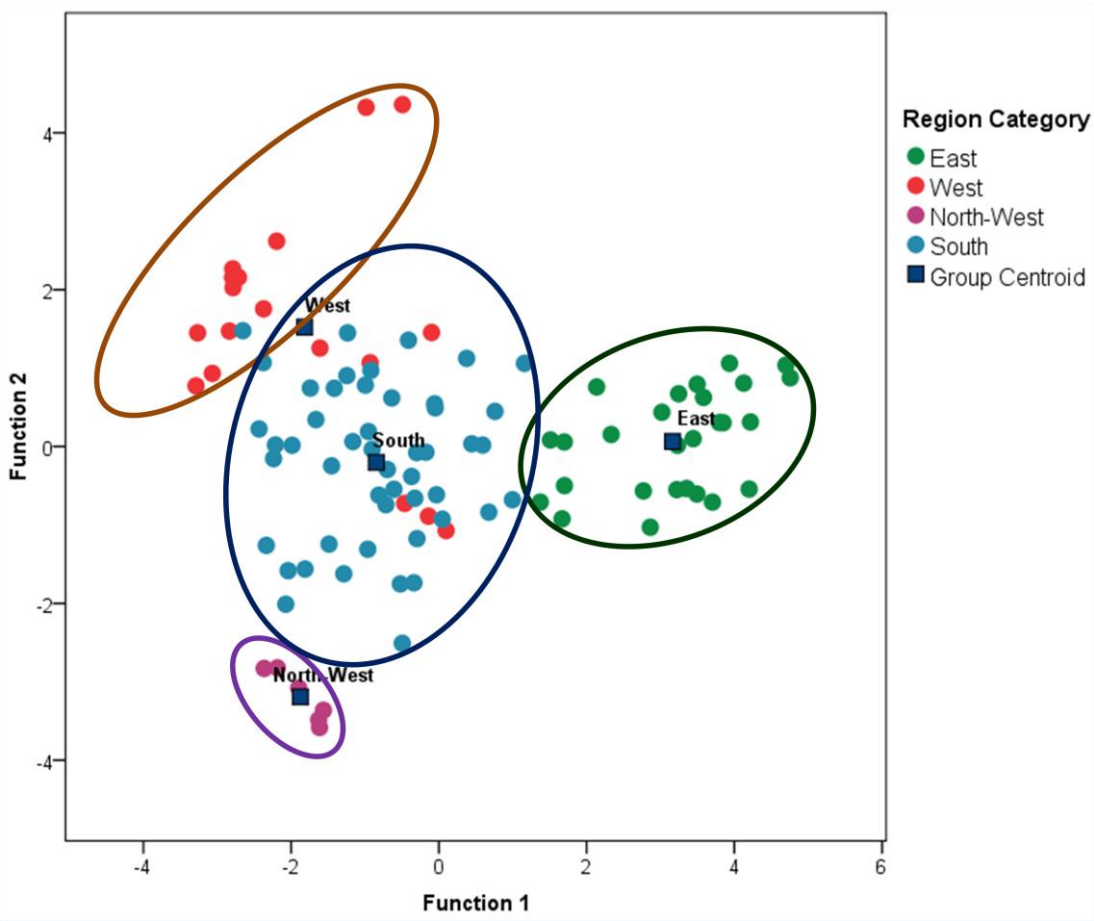



**Figure 5.2.1: Scatter plot of the first two canonical discriminant function scores of green coffee beans at regional level.**

### 5.2.2 Classification at Sub-Regional Level

Table 5.2.8 provides statistical evidence of significant differences between means of Sub-Regional categories for all variables with all 3-CQA, 5-CQA, 4-CQA, 5-pCoQA, 5-FQA, 3,4-

diCQA, 3,5-diCQA, 4,5-diCQA, 3,5-diCQA/4,5-diCQA, and 4,5-diCQA/43,4-diCQA producing very high value F's. For this study, Wilks' lambda is significant by the F test for all ten variables at p-value less than 0.05.

**Table 5.2.8: Tests of equality of group means table (sub-region)**

| | Wilks' Lambda | F | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| 3-CQA | .237 | 32.237 | 9 | 90 | .000 |
| 5-CQA | .581 | 7.207 | 9 | 90 | .000 |
| 4-CQA | .345 | 18.963 | 9 | 90 | .000 |
| 5-pCoQA | .786 | 2.723 | 9 | 90 | .007 |
| 5-FQA | .492 | 10.342 | 9 | 90 | .000 |
| 3,4-diCQA | .320 | 21.298 | 9 | 90 | .000 |
| 3,5-diCQA | .471 | 11.232 | 9 | 90 | .000 |
| 4,5-diCQA | .147 | 58.205 | 9 | 90 | .000 |
| 3,5-diCQA/4,5-diCQA | .249 | 30.195 | 9 | 90 | .000 |
| 4,5-diCQA/3,4-diCQA | .723 | 3.825 | 9 | 90 | .000 |

### 5.2.2.1 Summary of Canonical Discriminant Functions at Sub-region

This gives information on each of the discriminate functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus 1. In this case, ten groups were used, namely Harar, Jimma A, Jimma B, Kaffa, Wolega, Benishangul, Finote-Selam, Sidama SA, Sidama SB, and Yirgachefe. The canonical correlation measures the degree of association between the discriminant function and the predictors. In Table 6, a canonical correlation of 0.963 suggests that the model explains 92.74% (i.e.$0.963^2$ *100) of the variation in the grouping variable (Burns and Burns, 2008).

**Table 5.2.9: Eigenvalues table and canonical correlation (sub-region)**

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|---|---|---|---|---|
| 1 | 12.693[a] | 69.4 | 69.4 | .963 |
| 2 | 3.513[a] | 19.2 | 88.6 | .882 |
| 3 | .798[a] | 4.4 | 93.0 | .666 |
| 4 | .687[a] | 3.8 | 96.7 | .638 |
| 5 | .410[a] | 2.2 | 99.0 | .539 |
| 6 | .105[a] | .6 | 99.5 | .308 |
| 7 | .071[a] | .4 | 99.9 | .258 |
| 8 | .012[a] | .1 | 100.0 | .109 |

a. First 8 canonical discriminant functions were used in the analysis.

The result of this study shows a highly significant functions (sig. =0.000) and 0.3% unexplained (i.e. Wilks' lambda value of 0.003), which means using that combination of weights on the variables at sub-regional level leaves about 0.3% of the variance in the ten sub-regions (groups) unexplained.

**Table 5.2.10: Wilks' lambda table at sub-regional level**

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---|---|---|---|---|
| 1 through 8 | .003 | 518.225 | 72 | .000 |
| 2 through 8 | .043 | 282.706 | 56 | .000 |
| 3 through 8 | .195 | 147.087 | 42 | .000 |
| 4 through 8 | .351 | 94.306 | 30 | .000 |
| 5 through 8 | .592 | 47.219 | 20 | .001 |
| 6 through 8 | .835 | 16.268 | 12 | .179 |
| 7 through 8 | .922 | 7.283 | 6 | .295 |
| 8 | .988 | 1.082 | 2 | .582 |

From table below the result shows the variables 4,5-diCQA, 3,5-diCQA/4,5-diCQA, 3-CQA, and 3,5-diCQA were the strongest predictors for the discriminate functions, followed by 4,5-diCQA/3,4-diCQA, and 5-FQA compounds. These variables with large coefficients stand out as those that strongly predict the green coffee beans to geographical categories (groups) at sub-regional level.

The variables 5-CQA and 3,4-diCQA were excluded from the LDA model constructing at the sub-regional level. The forward stepwise variable selection method was used to select the most relevant variables (compounds) for the discrimination of coffees at sub-regional level.

**Table 5.2.11: Standardized canonical discriminant function coefficients table (sub-region)**

| | Function | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3-CQA | .321 | -.602 | .291 | -.027 | .090 | .367 | -.560 | .117 |
| 4-CQA | .263 | -.456 | .251 | -.079 | .169 | -.204 | .754 | -.271 |
| 5-pCoQA | -.085 | .073 | -.118 | .725 | -.313 | .616 | .092 | -.277 |
| 5-FQA | .185 | .126 | -.301 | .493 | .758 | .226 | -.058 | .331 |
| 3,5-diCQA | -.629 | -.396 | -.579 | -.610 | .458 | -.193 | -.482 | -1.148 |
| 4,5-diCQA | 1.341 | 1.086 | .642 | .014 | -.206 | .021 | .073 | .112 |
| 3,5-diCQA/4,5-diCQA | .605 | 1.060 | 1.540 | .510 | -.010 | .335 | .322 | .828 |
| 4,5-diCQA/3,4-diCQA | -.529 | -.082 | .041 | -.359 | .303 | .805 | .417 | .495 |

The structure matrix on table below shows the correlations of each variable with each of the discriminate function. The largest absolute coefficients for each discriminate function determine how each function is to be named. Similarly, as in component loadings, coefficients with absolute value $\geq 0.30$ are used as the cut-off between important and less important variables.

Accordingly, the variables 4,5-diCQA, 3-CQA, and 3,5-diCQA/4,5-diCQA are the most important predicators of the first, second and third functions, respectively. Similarly, the

variables 5-FQA, 4,5-diCQA/3,4-diCQA, 4-CQA, and 3,5-diCQA are the most important predicators of the remaining functions for the LDA model at sub-regional level.

**Table 5.2.12: Structure matrix table (sub-region)**

| | Function | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| 4,5-diCQA | .642* | .317 | -.321 | -.442 | -.072 | .377 | .078 | -.190 |
| 3-CQA | .375 | -.616* | .301 | -.070 | .010 | .364 | -.499 | .044 |
| 3,5-diCQA/4,5-diCQA | -.433 | .169 | .752* | .267 | .299 | -.144 | -.127 | -.145 |
| 5-FQA | .208 | .072 | -.327 | .408 | .809* | -.043 | -.017 | .147 |
| 4,5-diCQA/3,4-diCQA | -.036 | .110 | -.165 | -.591 | .056 | .700* | .304 | .156 |
| 3,4-diCQA[a] | .425 | .002 | -.107 | .445 | -.040 | -.617* | -.252 | -.263 |
| 5-pCoQA | -.016 | .026 | -.062 | .519 | -.300 | .589* | .136 | -.520 |
| 4-CQA | .290 | -.453 | .224 | .083 | .260 | -.079 | .686* | -.331 |
| 5-CQA[a] | -.078 | -.005 | -.042 | .122 | -.067 | .056 | .192* | -.018 |
| 3,5-diCQA | -.231 | .259 | .338 | -.140 | .477 | .136 | -.203 | -.676* |

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

From the canonical discriminant function coefficients table, we have unstandardized coefficients and are used to create the actual prediction equation, which are used to classify new cases. So, from the unstandardized coefficients table below, the prediction equation for discriminant functions at sub-regional level should be like this:

Di = Constant + B1*3-CQA + B2*4-CQA + B3*5-pCoQA + B4*5-FQA + B5*3,5-diCQA + B6*4,5-diCQA +B7*3,5-diCQA/4,5-diCQA + B8*4,5-diCQA/3,4-diCQA

**Table 5.2.13: Canonical Discriminant Function Coefficients table**

**Canonical Discriminant Function Coefficients**

| | Function | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3-CQA | 2.790 | -5.238 | 2.536 | -.236 | .781 | 3.193 | -4.870 | 1.014 |
| 4-CQA | 2.037 | -3.530 | 1.944 | -.614 | 1.309 | -1.580 | 5.837 | -2.099 |
| 5-pCoQA | -.402 | .344 | -.558 | 3.426 | -1.479 | 2.908 | .435 | -1.307 |
| 5-FQA | 1.108 | .757 | -1.804 | 2.960 | 4.552 | 1.357 | -.349 | 1.987 |
| 3,5-diCQA | -4.924 | -3.095 | -4.529 | -4.772 | 3.587 | -1.512 | -3.775 | -8.988 |
| 4,5-diCQA | 14.489 | 11.737 | 6.933 | .152 | -2.230 | .225 | .789 | 1.212 |
| 3,5-diCQA/4,5-diCQA | 5.799 | 10.164 | 14.760 | 4.884 | -.093 | 3.207 | 3.083 | 7.935 |
| 4,5-diCQA/3,4-diCQA | -3.507 | -.542 | .272 | -2.380 | 2.007 | 5.339 | 2.762 | 3.279 |
| (Constant) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |

Unstandardized coefficients

### 5.2.2.2 Classification Statistics at Sub-regional level

From the classification table below, result shows that all coffee samples from the Harar, Jimma A, Jimma B, Wolega, Benishangul, Finote-Selam, Sidama SA and Yirgachefe sub-regional groups were 100% correctly classified. On the other hand, all samples from two sub-regional groups were not 100% correctly classified. Among the 29 cases of Sidama SB, 28 (96.6%) of them were correctly classified, while only one sample was incorrectly classified into Finote-Selam. The lowest classification was obtained from Kaffa coffee samples in which 6 (60%) were correctly classified while 3 (30%) and 1 (10%) were incorrectly classified into Yirgachefe and Finoteselam, respectively.

In general, the classification results confirmed that 95.0% of the green coffee beans cases were classified correctly into the ten sub-regional areas of the country. In addition, using the leave-one-out cross-validation methods, in which each case is classified by the functions derived from all samples other than that sample, 88.0% of cross-validated grouped samples correctly classified at sub-regional level.

## Table 5.2.14: Sub-regional classification results

**Classification Results[b,c]**

| | | | Predicted Group Membership | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sub-Region Category | Harar | Jimma A | Jimma B | Kaffa | Wolega | Benishangul | Finote-Selam | Sidama SA | Sidama SB | Yirgachefe | Total |
| Original | Count | Harar | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 27 |
| | | Jimma A | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | | Jimma B | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | | Kaffa | 0 | 1 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 3 | 10 |
| | | Wolega | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| | | Benishangul | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 |
| | | Finote-Selam | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 |
| | | Sidama SA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 10 |
| | | Sidama SB | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 28 | 0 | 29 |
| | | Yirgachefe | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 10 |
| | % | Harar | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Jimma A | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Jimma B | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Kaffa | .0 | 10.0 | .0 | 60.0 | .0 | .0 | .0 | .0 | .0 | 30.0 | 100.0 |
| | | Wolega | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Benishangul | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Finote-Selam | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | 100.0 |
| | | Sidama SA | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | 100.0 |
| | | Sidama SB | .0 | .0 | .0 | .0 | .0 | .0 | 3.4 | .0 | 96.6 | .0 | 100.0 |
| | | Yirgachefe | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 | 100.0 |
| Cross-validated[a] | Count | Harar | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 27 |
| | | Jimma A | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | | Jimma B | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | | Kaffa | 0 | 2 | 0 | 3 | 0 | 0 | 0 | 2 | 0 | 3 | 10 |
| | | Wolega | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| | | Benishangul | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 3 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Finote-Selam | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 |
| | | Sidama SA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 10 |
| | | Sidama SB | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 26 | 1 | 29 |
| | | Yirgachefe | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 8 | 10 |
| | % | Harar | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Jimma A | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Jimma B | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Kaffa | .0 | 20.0 | .0 | 30.0 | .0 | .0 | .0 | 20.0 | .0 | 30.0 | 100.0 |
| | | Wolega | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Benishangul | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | .0 | 100.0 |
| | | Finote-Selam | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | .0 | 100.0 |
| | | Sidama SA | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 100.0 | .0 | .0 | 100.0 |
| | | Sidama SB | .0 | .0 | 3.4 | .0 | .0 | .0 | 3.4 | .0 | 89.7 | 3.4 | 100.0 |
| | | Yirgachefe | .0 | .0 | .0 | .0 | .0 | .0 | .0 | 20.0 | .0 | 80.0 | 100.0 |

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 95.0% of original grouped cases correctly classified.

c. 88.0% of cross-validated grouped cases correctly classified.

As it can be seen from the scatter plot below, the samples of Harar coffees are clearly separated to the positive parts of first function. Moreover, based on the coefficients of canonical discriminant function (Table 5.2.13), function 1 was highly influenced to the positive side by 4,5-diCQA. Hence, Harar coffees are separated by their higher content of 4,5-diCQA.

The samples of coffees from Benishangul and Finote-selam are clearly separated to the negative part of function 2. Based on the coefficients of canonical discriminant function (Table 5.2.13), the second function was highly influenced to the negative side by 3-CQA. Hence, Benishangul and Finote-selam coffees are separated by their higher content of 3-CQA.

Moreover, using the scatter plots and combination of those eight functions each other, two sub-region's coffee samples from South and three sub-regions' coffee samples from West were separated. Accordingly, Jimma A and Wollega coffees from West and Sidama SA coffees from South were clearly distinguished from others by their higher content of 3,5-

diCQA/4,5-diCQA concentrations ratios. Similarly, Yirgachefe coffees from South and Jimma B coffees from West are clearly distinguished from others by their higher content of 4,5-diCQA/3,4-diCQA concentrations ratios.

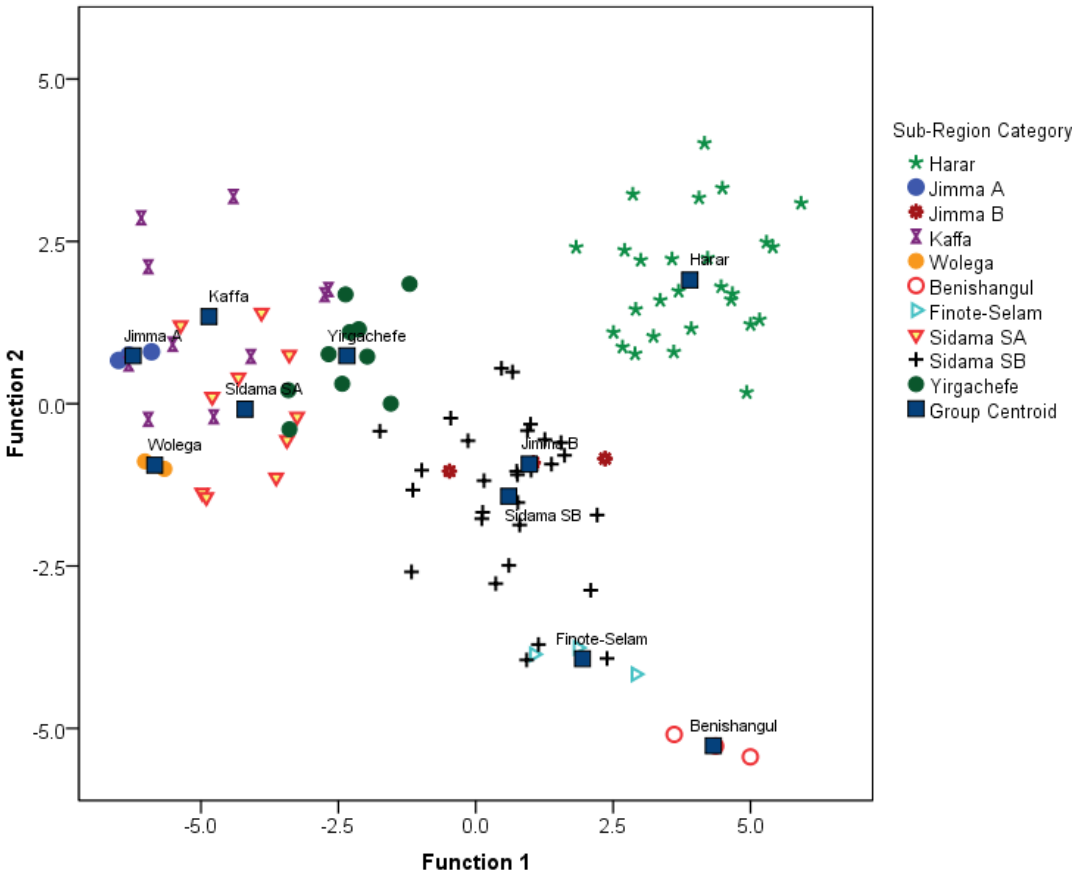



**Figure 5.2.2: Scatter plot of functions 1 and 2 for Green Coffee beans by Sub-Region**

## 5.3 Discussions

### 5.3.1 Principal component analysis (PCA)

In this study, green coffee beans were compared and classified by investigating if there are clusters relating to geographical regions of origin or coffee type. Principal component analysis can provide information as to how the coffee samples are related to one another, as well as enabling the selection of the most discriminating compounds useful for the differentiation of the various green coffee beans.

The dataset consisting of 10 variables corresponding to the compounds of the green coffee beans and 100 observations corresponding to samples of the green coffee beans were tested for the suitability of the data for the PCA model construction. Kaiser-Meyer-Olkin (KMO)

Measure of Sampling Adequacy and Bartlett's test of Sphericity (BTS), which determines whether the sampling was adequate to proceed with PCA (Maat, Zakaria, Nordin, & Meerah, 2011), were conducted on the green coffee beans dataset. KMO's acceptable level above 0.5 (Coakes et al., 2003; Hair et al., 2010; Kaiser, 1974), supporting the results by BTS, and item's communality that was above 0.3 (Tabachnick and Fidell, 2007) were used.

Accordingly, the results for two variables 5-pCoQA compound and 4,5-diCQA/3,4-diCQA concentration ratios were below the minimum recommended values. Moreover, the ANOVA tests of variance at regional level shows significant for all except the above two variables at p-value less than 0.05, while test of homogeneity of variances shows insignificant for only 5-pCoQA at regional level. However, both ANOVA tests of variance and test of homogeneity of variances show significant for all the 10 variables at sub-regional levels at p-value less than 0.05. Hence, the two variables 5-yCoQA and 4,5-diCQA/3,4-diCQA were removed and PCA was then conducted with 8 variables of green coffee beans with Varimax rotation.

The separation of the coffee samples by the first two components is illustrated by the **scores plot** (Figure 5.1.2), which indicates the presence of groups and some other important patterns in the data. Three of the coffee samples from the Northwest part of the country (Benishangul) are clearly very different from the other samples and lie far to the positive side of both PC2 and tend to cluster on the positive side of PC1, two of the samples revealed as outliers. However, these samples were retained since their removal resulted in a poorer model as reflected by the new model parameters.

The coffees from the East are different from the other samples, since they form a distinct group to the positive side of PC1 and negative side of PC2 (or tend to cluster on negative side of PC2). Samples from the West, with the exception of the three Jimma B samples, lie to the negative side of PC1 and positive side of PC2. Coffee samples from the southern part of Ethiopia do not form a unique cluster, but instead they have an extended distribution on the plane formed by the two principal components. However, the distribution revealed some important trends with respect to coffee types. All of the Sidama SA and Yirgachefe samples scatter on the negative side of PC1. On the other hand, almost all the Sidama SB samples tend to scatter on the positive side of PC1.

From the loadings plot (Figure 5.1.1) it is evident that coffees from East are differentiated mainly due to 4,5-diCQA, which lies far to the negative side of PC2. In contrast, coffees from

West are characterized by 3,5-diCQA to 4,5-diCQA concentration ratios (which lies far to the negative side of PC1 and positive side of PC2) and 3,5-diCQA (which lies far to the negative side of PC1), while coffee samples from Northwest differentiated mainly due to 3-CQA (which lies far to the positive side of PC1).

### 5.3.2 Quantitative analysis of CGAs

All of the eight variables i.e. 3-CQA, 5-CQA, 4- CQA, 5-FQA, 3,4-diCQA, 3,5-diCQA, 4,5-diCQA, and 3,5-diCQA to 4,5-diCQA concentration ratios, contributed significantly towards the differentiation of samples as revealed by the relatively high PCA loadings (Figure 5.1.1) and PCA component matrix (Table 5.1.7).

Univariate statistical tests ($\alpha = 0.05$) to ascertain the effect of region and sub-region on the levels of the various CGAs was carried out (Appendix A). Accordingly, the four regional coffee samples have been found to contain similar levels of total CGAs with no statistically significant difference among each other. However, the amounts of the various individual CGAs have been found to differ significantly among the regions. The exception to this finding are 5-pCoQA and 4,5-diCQA to 3,4-diCQA concentration ratios, which appear to be randomly distributed with no significant difference found between the regions studied. In line with this, Mehari, B. et al. (2016) found, with the exception of 5-pCoQA, similar levels of total CGAs with no statistically significant difference in Ethiopian green coffee beans from different four studied regional categories. Equally, Joet et al. (2010) also found no significant difference in the total concentration of CGAs in Arabica green coffee beans from different locations in Reunion Island. However, both studies have observed variations in the concentration of individual CGAs depending on the growing location of coffee beans.

Among the determined CGAs, 5-CQA was the most abundant in the green coffee beans. The average concentration of 5-CQA (average 31.5 mg/g) is 52% of the total 8 identified CGAs, which is within the range of 46– 59% studied by Mehari, B. et al. (2016). Coffees from the western part of Ethiopia contain higher concentrations of 5-CQA (average 33.8 mg/g), while East (Harar) coffees contain the lowest levels (average 29.0 mg/g) of 5-CQA. The lower concentrations of 5-CQA in Harar coffees may contribute towards the much appreciated cup quality characteristics of these coffees *(*Mehari, B. et al., 2016).

Green coffee beans from the Northwest region category contain significantly higher amounts of both 3-CQA (average 6.2 mg/g) and 4-CQA (average 8.4 mg/g) followed by coffees from East 3-CQA (average 3.9 mg/g) and 4-CQA (average 6.8 mg/g) while coffees from West show the lowest levels of average 2.9 mg/g and 5.9 mg/g, respectively. *G*reen coffee beans from East contain significantly higher amounts of both 4,5-diCQA (average 4.9 mg/g) and 3,4-diCQA (average 2.9 mg/g), while coffee beans from the West region show the lowest levels of 4,5-diCQA (average 2.6 mg/g) and 3,4-diCQA (average 1.6 mg/g). East coffees contain significantly higher level of 3,4-diCQA (average 2.9 mg/g) than West coffees (average 1.6 mg/g). Furthermore, green coffee beans from East contains higher amounts of 5-FQA (average 5.7 mg/g) followed by coffees from Southern Ethiopia (average 5.3 mg/g) while coffees from West contain the lowest level (average 4.5 mg/g) of 5-FQA (Table 5.3.1).

On the other hand, West coffees contain significantly higher amounts of 3,5-diCQA (average 8.6 mg/g) followed by coffees from the Southern Ethiopia (average 6.8 mg/g) while coffees from Northwest region show the lowest level (average 5.1 mg/g) of 3,5-diCQA. Moreover, coffees from West show significantly higher amounts of 3,5-diCQA/4,5-diCQA concentration ratio (average 3.8 mg/g) followed by coffees from the Southern Ethiopia (average 2.0 mg/g) while coffees from East and Northwest region both show the lowest level of average 1.3 mg/g (Table 5.3.1).

**Table 5.3.1: Average CGA concentration of green coffee beans by region**

| Region | 3-CQA | 5-CQA | 4-CQA | 5-pCoQA | 5-FQA | 3,4-diCQA | 3,5-diCQA | 4,5-diCQA | 3,5-diCQA/4,5-diCQA | 4,5-diCQA/3,4-diCQA | Total (10) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| East | 3.9 | 29.0 | 6.8 | 0.7 | 5.7 | 2.9 | 6.3 | 4.9 | 1.3 | 1.7 | 63.2 |
| West | 2.9 | 33.8 | 5.9 | 0.7 | 4.5 | 1.6 | 8.6 | 2.6 | 3.8 | 1.7 | 66.0 |
| NW | 6.2 | 32.8 | 8.4 | 0.7 | 4.7 | 2.6 | 5.0 | 3.9 | 1.3 | 1.5 | 67.0 |
| South | 3.7 | 31.8 | 6.6 | 0.7 | 5.3 | 2.1 | 6.8 | 3.5 | 2.0 | 1.8 | 64.2 |
| **Mean** | **3.7** | **31.5** | **6.6** | **0.7** | **5.2** | **2.3** | **6.9** | **3.8** | **2.1** | **1.7** | **64.4** |
| **%** | 6.2 | 51.9 | 11.0 | 1.1 | 8.6 | 3.7 | 11.4 | 6.2 | | | **100.0** |
| **%** | 5.8 | 48.8 | 10.3 | 1.1 | 8.1 | 3.5 | 10.7 | 5.8 | 3.2 | 2.7 | **100.0** |

At sub-regional level, green coffee beans samples from Benishangul contain significantly higher amounts of both 3-CQA (average 6.5 mg/g) and 4-CQA (average 9.3 mg/g) followed

by sample coffees from Finoteselam 3-CQA (average 5.9 mg/g) and 4-CQA (average 7.5 mg/g. *G*reen coffee beans from Hara contain significantly higher amounts of 4,5-diCQA (average 4.9 mg/g) followed by green coffees samples from Jimma B and Benishangul both (average 4.1 mg/g). Moreover, coffee samples from Harar and Sidama SB both contain significantly higher amounts of 5-FQA (average 5.7 mg/g) (Table 5.3.2).

On the other hand, green coffee samples from Kaffa contain significantly higher amounts of 3,5-diCQA (average 9.8 mg/g) followed by coffees from Jimma A (average 8.9 mg/g), Yirgachefe coffees (average 8.1 mg/g), Sidama SA coffees (average 7.1 mg/g) and coffees from Wollega (average 7.0 mg/g). Moreover, green coffees from Jimma A show significantly higher amounts of 3,5-diCQA/4,5-diCQA concentration ratio (average 4.5 mg/g). Green coffees from Kaffa show the second higher amount of this concentration ratio (average 4.3 mg/g), followed by coffees from Wollega (average 3.3 mg/g) and Sidama SA (average 2.6 mg/g) while coffees from Jimma B ((average 1.4 mg/g), Harar ((average 1.3 mg/g) and Benishangul ((average 1.2 mg/g) show the 3 lowest amount of this concentration ratio. Similarly, green coffee beans from Yirgachefe show significantly higher amounts of 4,5-diCQA/3,4-diCQA concentration ratio (average 2.4 mg/g) and coffees from Jimma B show the second higher amounts of this concentration ratio (average 2.0 mg/g) (Table 5.3.2).

**Table 5.3.2: Average CGA concentration of the of green coffee beans by sub-region**

| Sub-Region | 3-CQA | 5-CQA | 4-CQA | 5-pCoQA | 5-FQA | 3,4-diCQA | 3,5-diCQA | 4,5-diCQA | 3,5-diCQA/4,5-diCQA | 4,5-diCQA/3,4-diCQA | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Harar | 3.9 | 29.0 | 6.8 | 0.7 | 5.7 | 2.9 | 6.3 | 4.9 | 1.3 | 1.7 | 63.2 |
| Jimma A | 2.5 | 35.3 | 5.6 | 0.9 | 4.5 | 1.2 | 8.8 | 2.0 | 4.5 | 1.6 | 67.0 |
| Jimma B | 4.2 | 32.0 | 6.9 | 0.5 | 4.2 | 2.1 | 5.6 | 4.1 | 1.4 | 2.0 | 62.9 |
| Kaffa | 2.6 | 34.0 | 5.7 | 0.6 | 4.6 | 1.6 | 9.8 | 2.5 | 4.3 | 1.7 | 67.5 |
| Wollega | 2.9 | 33.0 | 5.8 | 0.8 | 4.1 | 1.3 | 7.0 | 2.1 | 3.3 | 1.6 | 61.9 |
| Benishangul | 6.5 | 33.0 | 9.3 | 0.6 | 4.6 | 2.6 | 4.8 | 4.1 | 1.2 | 1.6 | 68.3 |
| Finoteselam | 5.9 | 32.7 | 7.5 | 0.7 | 4.7 | 2.6 | 5.2 | 3.7 | 1.4 | 1.4 | 65.8 |
| Sidama SA | 2.7 | 31.0 | 5.7 | 0.7 | 4.5 | 1.7 | 7.1 | 2.8 | 2.6 | 1.7 | 60.5 |
| Sidama SB | 4.2 | 32.0 | 7.2 | 0.7 | 5.7 | 2.4 | 6.3 | 3.7 | 1.7 | 1.6 | 65.5 |
| Yirgachefe | 3.1 | 32.0 | 5.9 | 0.6 | 4.8 | 1.6 | 8.1 | 3.7 | 2.2 | 2.4 | 64.4 |
| **Total** | **3.7** | **31.5** | **6.6** | **0.7** | **5.2** | **2.3** | **6.9** | **3.8** | **2.1** | **1.7** | **64.4** |
| **Percent (%)** | **5.8** | **48.8** | **10.3** | **1.1** | **8.1** | **3.5** | **10.7** | **5.8** | **3.2** | **2.7** | **100.0** |

As previously illustrated by using PCA, coffees from East (Hara sub-regional) are best discriminated from the other regional (as well as other sub-regional) coffee beans by 4,5-diCQA. These coffees contain significantly higher amount of 4,5-diCQA (average 4.9 mg/g) than coffee beans from the other regions (average 2.6–3.9 mg/g). The individual samples from this region have more than 4.5 mg/g 4,5-diCQA, which is higher than the concentration in any other sample (Table 5.3.2).

Coffee samples from Northwest are distinguished from those of other region coffees by higher contents of 3-CQA. All of the individual coffee samples from this particular region contained more than 5.5 mg/g of 3-CQA (Table 5.3.2). Results of one-way ANOVA ($\alpha$ = 0.05) also indicated that the mean 3-CQA content of Northwest coffees (6.2 mg/g) differ significantly from the other regional coffees, which were in the range of 2.9–3.9 mg/g.

The PCA scores plot (Figure 5.2) indicates that green coffee beans from West can be characterized by their higher 3,5-diCQA/4,5-diCQA concentration ratio, i.e. higher content of 3,5-diCQA but smaller content of 4,5-diCQA. With the exception of Jimma B coffees, this concentration ratio appears to be suitable to distinguish all of West coffees. The PCA superposition plot (Figure 5.1.3) of 3,5-diCQA/4,5-diCQA concentration ratio was able to separate four sub-regional coffee samples, i.e. Jimma A, Kaffa Wollega (all from West) and Sidama A( from South), from the other types of green coffee beans. Results of one-way ANOVA ($\alpha$ = 0.05) indicated that the mean 3,5-diCQA/4,5-diCQA concentration ratio (3.7), calculated from the four coffee types, differ significantly from the other sub-regional coffees, which were in the range of 1.2–2.2. On the other hand, Jimma B coffees, which show the third lowest amount ((average 1.4 mg/g), cannot be differentiated from the other sub-regional coffees based on the 3,5-diCQA/4,5-diCQA concentration ratio (Table 5.3.2).

Yirgachefe coffees from South and Jimma B from West could be distinctly differentiated from the other coffee samples by using the 4,5-diCQA/3,4-diCQA concentration ratio (Table 5.2). Results of one-way ANOVA ($\alpha$ = 0.05) indicated that the mean 4,5-diCQA/3,4-diCQA concentration ratio of Yirgachefe and Jimma B coffees (2.2) differs significantly from those of the other sub-regional coffees, which were in the range of 1.4–1.7. The PCA superposition plot (Figure 5.1.3) also shows that 4,5-diCQA/3,4-diCQA concentration ratio was in the direction of Yirgachefe and Jimma B coffees.

Based on the distribution of coffees from Sidama SB, the PCA score plot couldn't identify any compound that strictly distinguishes these coffees from the other sub-regional coffee samples. However, from the PCA superposition plot (Figure 5.1.3), majority of the green coffees from Sidama SB were distinguished by 5-FCQ compound.

The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived from all samples other than that sample in the entire dataset. Accordingly, 93% of cross-validated groups of samples were correctly classified.

### 5.3.3 Discriminant analysis at regional level

PCA, unsupervised method, revealed potential clustering of coffees according to their origin. Consequently, a discriminant analysis model that can be used for the authentication of the geographical origin of the green coffee beans was constructed from the concentration of the eight determined CGAs and two concentration ratios. To achieve this aim, linear discriminant analysis (LDA) modeling was applied. Firstly, the classification of the coffee beans at a regional level, i.e. East, Northwest, West and South, was studied. The forward stepwise variable selection method was used to select the most relevant variables (CGAs) for the discrimination of regional coffees. Consequently, the variables 4-CQA, 3,4-diCQA, 5-yCoQa and 4,5-diCQA/3,4-diCQA concentration ratio were excluded from the LDA modeling **at the regional level**. Three canonical discriminant functions were computed. The magnitude of Wilks' $\lambda$, which reflects the proportion of the variance in the dataset that is not accounted for by the model, encompassing all three functions of the model, was 0.056 (i.e. 5.6%), reflecting the good discriminating ability of the LDA model. The canonical correlation, which measures the degree of association between the discriminant function and the predictors, shows 0.895 suggesting that the model explains 80.10% (i.e.$0.895^2$ *100) of the variation in the grouping variable (Burns and Burns, 2008).

The result of structure matrix of the discriminant functions (Table 9) shows that 4,5-diCQA compound is the most important predicator of the first function. Similarly, the compound 3-CQA is the most important predicator of the second function followed by 3,5-diCQA/4,5-diCQA concentration ratio, while 5-FQA compound is the highest predicator of the third function of the LDA model at regional level. Examination of the coefficients of the discriminant functions (Table 5.2.6) indicates that the compounds 4,5-diCQA, 3-CQA, and

3,5-diCQA/4,5-diCQA concentration ratio are the highest contributors to the first and second functions, and hence, are the most discriminating compounds among the green coffee beans collected from the four regions studied. The scatter plot presented in Figure 5.2.1 illustrates the efficiency of the CGA composition for the discrimination of the green coffee beans from the four regions studied. The first discriminant function allowed excellent separation of East (Harar) coffees from the other region samples. This function was highly influenced to the positive side by 4,5-diCQA followed by 3,5-diCQA/4,5-diCQA (Table 5.2.6). Hence, East (Harar) coffees are varied mainly from the higher contents of 4,5-diCQA. Similarly, the second discriminant function allowed separates Northwest coffees from the other region samples. This function was highly influenced to the negative side by 3-CQA (Table 5.2.6). Hence, Northwest coffees are varied mainly from their higher content of 3-CQA. Coffee samples from West and South tend to show a similar CGA profile that is evident from their partial overlap on the PCA scores plot (Figure 5.2.1). However, they are significantly separated from one another by the combined effect of the first, second, and third discriminant functions.

From the summary of classification results, green coffee samples from East and Northwest were 100% correctly classified, whereas coffees from South and West were 98% and 66.7% correctly classified, respectively. The overall percentage of the sample set correctly classified was 93%. The misclassified samples were six samples from West that were incorrectly classified as South and only one sample from South that was incorrectly classified as West. This overall proportion of correct classification obtained in this study (93%) is better than that (92%) correct classification obtained by Mehari, B. et al. (2016) and (83%) correct classification obtained by Bertrand et al. (2008) as cited in (Mehari, B. et al., 2016) following linear discriminant analysis of green Arabica coffee beans from three Colombian locations, based on their CGA contents.

The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived from all samples other than that sample in the entire dataset. Accordingly, 93% of cross-validated groups of samples were correctly classified.

The reliability of the LDA model was assessed in terms of its recognition and prediction abilities. For this, the entire sample set was divided into a training set and testing set, which was as an external validation set. The testing set consisted of 30 (30%) randomly selected

samples, while, the remaining 70 (70%) samples were used as training set to construct the LDA model. This proportion was also used for the Northwest coffee samples, and hence, 4 samples were included in training set and the remaining two were included in testing set. Accordingly, the recognition ability of the model, calculated as the percentage of the members of the training set that are correctly classified, was 94%. Moreover, the prediction ability of the LDA model, calculated as the percentage of the members of the external validation set correctly classified by using the model developed in the training step, was 92.2% at the regional level.

### 5.3.4 Discriminant analysis at sub-regional level

The classification of the green coffee beans was conducted for the ten sub-regions. Again the forward stepwise variable selection method was used to select the most relevant variables (CGAs) for the discrimination of sub-regional coffees. Consequently, the variables 5-CQA and 3,4-diCQA were excluded from the LDA modeling **at the sub-regional level**. Eight canonical discriminant functions were computed. The magnitude of Wilks' λ, which reflects the proportion of the variance in the dataset that is not accounted for by the model, encompassing all functions of the model, was 0.003 (i.e. 0.3%), reflecting the higher discriminating ability of the LDA model. The canonical correlation, which measures the degree of association between the discriminant function and the predictors, shows 0.895 suggesting that the model explains 92.74% (i.e.$0.963^2$ *100) of the variation in the grouping variable (Burns and Burns, 2008).

The result of coefficients of the discriminant functions (Table 5.2.13) at sub-regional level indicates that the compounds 4,5-diCQA, 3-CQA, and 3,5-diCQA/4,5-diCQA concentration ratio are the highest contributors to the first and second functions, and hence, are the most discriminating compounds among the green coffee beans collected from the four regions studied. Moreover, the scatter plot presented in Figure 5.1.6 illustrates the efficiency of the CGA composition for the discrimination of the green coffee beans from the ten sub-regions. The first discriminant function allowed excellent separation of Harar coffees from the other region samples. This function was highly influenced to the positive side by 4,5-diCQA (Table 5.2.13). Hence, East (Harar) coffees are varied mainly from the higher contents of 4,5-diCQA. Similarly, the second discriminant function allowed excellent separation of Benishangul and Finoteselam sample coffees from the other samples. This function was

highly influenced to the negative side by 3-CQA (Table 5.2.13). Hence, Benishangul and Finoteselam coffees are separated mainly from their higher content of 3-CQA.

Moreover, by combining the eight discriminant functions, with the exception of coffees from Kaffa and Sidama SB, the remaining coffee samples were clearly separated from one another. Accordingly, the positive side of the third discriminant function separated Jimma A sample coffees from the other samples. This function was highly influenced to the positive side by 3,5-diCQA/4,5-diCQA concentration ratio. The positive side of this function also separated most the coffees from Kaffa by 3,5-diCQA/4,5-diCQA concentration ratio. Similarly, the positive side of the fourth discriminant function separated Wollega and Sidama SA coffees from the other samples. This function was highly influenced to the positive side by 3,5-diCQA/4,5-diCQA concentration ratio (Table 5.2.13). The positive side of the fifth discriminant function separated most of Sidama SB coffees from the other samples. This function was highly influenced to the positive side by 5-FQA. The remaining two sub-regional coffees, coffees from Yirgachefe and Jimma B were separated on the positive side of the sixth discriminant function, which was highly influenced to the positive side by 4,5-diCQA/3,4-diCQA concentration ratio (Table 5.2.13).

From the summary of classification results (Table 5.2.14), green coffee samples from Harar, Jimma A, Jimma B, Wollega, Benishangul, Finoteselam, Sidama SA and Yirgachefe were 100% correctly classified, whereas coffees from Sidama SB and Kaffa were 96.6% and 60% correctly classified, respectively. The overall percentage of the sample set correctly classified was 95%. The misclassified samples were three samples from Kaffa that were incorrectly classified as Yirgachefe, one sample from Kaffa incorrectly classified in Jimma A and one sample from Sidama SB that was incorrectly classified as Finoteselam.

This overall proportion of correct classification obtained in this study (95%) is better than that (89%) obtained by Mehari, B. et al. (2016), combining the ten sub-regions into seven, i.e. Harar coffees, Jimma (Jimma A and Jimma B in one class), Kaffa, Wollega, Sidama (Sidama SA and Sidama SB in one class), Yirgachefe, and Northwest coffees (Benishangul and Finoteselam in one class). The validity of the LDA model was assessed by means of leave-one-out cross-validation, in which each sample was classified by the discriminant functions derived from all samples other than that sample in the entire dataset. Accordingly, 88% of cross-validated groups of samples were correctly classified.

The reliability of the LDA model was assessed in terms of its recognition and prediction abilities. For this, the entire sample set was divided into a Training set and Testing set. The testing set consisted of 30 randomly selected samples, while, the remaining 70 samples were used as training set to construct the LDA model. Due to the sample size, all of samples of Jimma A (3), Jimma B (3), Wollega (2), Benishangul (3), and Finoteselam (3) were used exclusively in the testing set. Accordingly, the recognition ability of the model, calculated as the percentage of the members of the training set that are correctly classified, was 94.3% at sub-regional level. Moreover, the prediction ability of the LDA model, calculated as the percentage of the members of the testing set correctly classified by using the model developed in the training step, was 93.3%, at the sub-regional level.

# CHAPTER

# 6 CONCLUSIONS AND RECOMMENDATIONS

## 6.1 Conclusions

In the present work, dimensionality reduction and classification algorithms, which address the authentication of coffee samples according to their geographical origin, were easily developed using Principal component analysis (PCA) and Linear discriminant analysis (LDA). The study used improved PCA to identify the most important discriminant variables (CGAs), aiming to assist the concerned bodies to better understand the authentication of coffees based on their geographical origin. As a tool for eliminating information overlaps, the improved PCA in this study cannot only retain more dispersion degree information in the original dataset, but also has a wider range of applicability. Exploratory analysis by PCA and LDA using the obtained coffee samples showed, in general, good discrimination capabilities among the different regional coffees varieties. Accordingly, 3-caffeoylquinic acid (3-CQA), 4,5-dicaffeoylquinic acid (4,5-diCQA), 3,5-dicaffeoylquinic acid (3,5-diCQA) to 4,5-dicaffeoylquinic acid (4,5-diCQA) concentration ratio and 4,5-dicaffeoylquinic acid (4,5-diCQA) to 3,4-dicaffeoylquinic acid (3,4-diCQA) concentration ratio were identified as the most discriminating compounds for the authentication of the various regional green coffee beans. Among these, 3-caffeoylquinic acid (3-CQA) and 4,5-dicaffeoylquinic acid (4,5-diCQA) were selected as suitable discriminant marker compounds for green coffee beans originating from Northwest and East growing regions, respectively. In contrast, no specific CGA could be identified to differentiate green coffee beans from the South and West growing at regional levels. However, the concentration ratios of two different CGAs were found to be suitable for the discrimination of green coffee beans at the sub-regional levels. Accordingly, the best discrimination was achieved with coffee beans from Jimma A, Wollega, and Sidama SA, which were distinguished by the 3,5-diCQA to 4,5-diCQA concentration ratio. For coffee beans from Yirgachefe and Jimma B, the 4,5-diCQA to 3,4-diCQA concentration ratio was found appropriate to differentiate these two sub-regional coffee beans from the other coffee varieties. In general, the results of DA were in line with the PCA results, indicating that the LDA model was able to classify the coffee beans accurately based on the their regional as well as sub-regional geographical origin. The recognition and prediction abilities of the LDA model were 94% and 92.4%, respectively, at the regional level and 94.3% and 93.3%,

respectively, at the sub-regional level and hence, best discrimination of green coffee beans was achieved both at regional and sub-regional.

## 6.2 Recommendations and Future Research Directions

Based on the findings of the study and conclusions made, the following possible recommendations are drawn.

As the findings of the study, the most important discriminant compounds (CGAs) that distinguished the coffee beans based on their geographical origin were identified. Hence, this study recommends that the government and concerned bodies should use it and work to prevent fraudulent practices and address the authentication of coffees based on their geographical origin.

The present study was conducted to identify the most important discriminant factors that classify the green coffee beans using improved PCA and LDA. Therefore, there is scope for other researchers to use others methods to develop better classification algorithms.

Similarly, future researcher may use more sample size for each regional and sub-regional categories in order to get more and better results. Besides that, since there exists an overlapping among few categories of coffee beans, the future researchers may reinvestigate the regional and sub-regional categories, particularly, the South categories.

## REFERENCES


A. Tharwat, A.E. Hassanien and B.E. Elnaghi, A ba-based algorithm for parameter optimization of support vector machine, *Pattern Recognition Letters* (2016).

Adnan Ullah, Usman Qamar, Farhan Hassan Khan, and Saba Bashir (2017). Dimensionality Reduction Approaches and Evolving Challenges in High Dimensional Data. **Conference Paper**, October 2017. DOI: 10.1145/3109761.3158407

Arunasakthi. K and Kamatchipriya. L(2014), _A Review On Linear And Non-Linear Dimensionality Reduction Techniques', Machine Learning And Applications: An Int. J. (Mlaij), Vol.1, No.1, Pp.65-76.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bay, S. D., Kibler, D., Pazzani, M. J., and Smyth, P. (2000). The UCI KDD archive of large data sets for data mining research and experimentation. ACM SIGKDD Explorations Newsletter, Volume 2,Issue 2,pp. 81-85.

Bertrand B, Villarreal D, Laffargue A, Posada H, Lashermes P, Dussert S (2008) Comparison of the effectiveness of fatty acids, chlorogenic acids, and elements for the chemometric discrimination of coffee (*Coffea arabica* L.) varieties and growing origins. J Agric Food Chem 56:2273–2280

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Information science and statistics. New York : Springer.

Burns, R.P., & Burns, R. (2008). *Business research methods and statistics using SPSS*. Sage.

Cai,W.; Dou, L.M.; Si, G.Y.; Cao, A.Y.; He, J.; Liu, S. (2016). A principal component analysis/fuzzy comprehensive evaluation model for coal burst liability assessment. Int. J. Rock Mech. Min. Sci. **2016**, 81, 62–69. [CrossRef]

Cattell, R. B. 1966. The scree test for the number of factors. Multivariate Behavioral Research 1: 245–276.

Chandrasekhar, T., Thangavel, K., and Elayaraja, E. (2012). Effective Clustering Algorithms for Gene Expression Data, arXiv preprint arXiv:1201.4914, International Journal of Computer Applications, Volume 32, No.4.

Charu C. Aggarwal (2014). Data Classification, Algorithms and Applications. Chapman and Hall/CRC 2014, Pages 37–64 Print ISBN: 978-1-4665-8674-1 eBook ISBN: 978-1-4665-8675-8

Chen, C.J., & Hung, S.W. (2010). To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities. *Information & Management, 47*(4), 226-236.

Coussement, A.; Isaac, B.J.; Gicquel, O.; Parente, A. (2016). Assessment of di_erent chemistry reduction methods based on principal component analysis: Comparison of

the MG-PCA and score-PCA approaches. Combust. Flame **2016**, 168, 83–97. [CrossRef]

Cunningham JP, Ghahramani Z. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*. 2015; 16:2859–2900.

Dash, R., Dash, R., and Mishra,D.(2010).A Hybridized Rough-PCA Approach of Attribute Reduction for High Dimensional Data Set, European Journal of Scientific Research, Volume 44 ,Issue 1, pp. 29.

Doorsamy, W.; Cronje, W.A. (2015). A Method for Fault Detection on Synchronous Generators Using Modified Principal Component Analysis. In Proceedings of the 2015 IEEE International Conference on Industrial Technology (ICIT), Seville, Spain, 17–19 March 2015; pp. 586–591.

Dr. H.B.Kekre, Sudeep D. Thepade and Akshay Maloo (2010), "CBIR Feature Vector Dimension Reduction with Eigenvectors of Covariance Matrix using Row, Column and Diagonal Mean Sequences", Int. J. of Computer Applications (0975 – 8887),Vol. 3, No.12.

Duras, T. (2019). Applications of Common Principal Components in Multivariate and High-Dimensional Analysis. *Doctoral Thesis in Statistics*, JIBS Dissertation Series No.131, Jonkoping International Business School.

F Kurniawan, I W Budiastra, Sutrisno, S Widyotomo. Classification of Arabica Java Coffee Beans Based on Their Origin using NIR Spectroscopy. IOP Conf. Series: Earth and Environmental Science 309 (2019) 012006.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks – New Delhi: Sage publications.

Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*. 1936; 7 (2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Geng, L., and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey, ACM Computing Surveys (CSUR), Volume 38, Issue 3, pp. 9.

Gupta, H., Agrawal, A. K., Pruthi, T., Shekhar, C., and Chellappa, R. (2002). An experimental evaluation of linear and kernel-based methods for face recognition, Sixth IEEE Workshop on Computer Vision, pp. 13-18.

Hair, J. F., Black, W. C., & Babin, B. J. (2010). RE Anderson Multivariate data analysis: A global perspective. New Jersey, Pearson Prentice Hall,).

Hao, R.X.; Li, S.M.; Li, J.B.; Zhang, Q.K.; Liu, F. (2013). Water Quality Assessment forWastewater Reclamation Using Principal Component Analysis. J. Environ. Inform. **2013**, 21, 45–54. [CrossRef]

Hapsari and Syamsuryadi (2019). Weather Classification Based on Hybrid Cloud Image Using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA); *J. Phys.: Conf. Ser.* **1167** 012064; www.doi:10.1088/1742-6596/1167/1/012064

Hoi, S. C., Lyu, M. R., and Chang, E. Y. (2006). Learning the unified kernel machines for classification, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 187-196.

Holmes S, Huber W. Modern Statistics for Modern Biology. Cambridge, UK: Cambridge University Press; 2019 [cited 2019 May 30]. Available from: https://www.huber.embl.de/msmb/.

Horn,J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Hosseini, H.M.; Kaneko, S (2011). Dynamic sustainability assessment of countries at the macro level: A principal component analysis. Ecol. Indic. **2011**, 11, 811–823. [CrossRef]

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Howes, Lee; Thomas, David (2008). *GPU Gems 3 - Efficient Random Number Generation and Application Using CUDA* . Pearson Education, Inc. ISBN 978-0-321-51526-1.

Huan Wang, S.C. Yan, D.Xu, X.O. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In IEEE Conference on Computer Vision and Pattern Recognition, 2007, pages 17–22, 2007.

Ivosev, G., Burton, L., and Bonner, R. (2008). Dimensionality reduction and visualization in principal component analysis, Analytical chemistry, Volume 80, Issue 13, pp. 4933-4944.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York.

J. Ye, R. Janardan and Q. Li, Two-dimensional linear discriminant analysis, in: *Proceedings of 17th Advances in Neural Information Processing Systems (NIPS)*, 2004, pp. 1569–1576.

Jayalakshmi, T., and Santhakumaran, A. (2009). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, Volume 3, Issue 1, pp. 1793-8201.

Jian Vora, (2019). Principal Component Analysis Stochastic Optimization. *arXiv*:1901.01798v1 [cs.LG] 7 Jan 2019

Joet, T., Laffargue, A., Descroix, F., Doulbeau, S., Bertrand, B., de Kochko, A., Dussert, S., 2010. Influence of environmental factors, wet processing and their interactions on the biochemical composition of green Arabica coffee beans. Food Chem. 118, 693–701.

Johnson RA and Wichern DW: *Applied Multivariate Statistical Analysis*. 6th edition. Pearson Prentice Hall; 2008: 430 – 470

Jolliffe I.T. (2002). *Prinicipal Component Analysis, 2nd Edition*, Springer series in statistics 2002, page 1-3.

Jolliffe IT, Cadima J. 2016. *Principal component analysis: a review and recent developments*. Phil. Trans. R. Soc. A 374: 20150202. http://dx.doi.org/10.1098/rsta.2015.0202

Jolliffe, I. (2014). Principal component analysis. *Wiley StatsRef: Statistics Reference Online*.

Julie M. David, Kannan Balakrishnan, (2012), Attribute Reduction and Missing Value Imputing with ANN: Prediction of Learning Disabilities, International Journal of Neural Computing & Applications, Springer-Verlag London Limited, DOI: 10.1007/s00521-011-0619, Vol. 21, Issue 7, pp 1757-1763

Julie M. David, Kannan Balakrishnan, (2014), Learning Disability Prediction Tool using ANN and ANFIS, International Journal of Soft Computing, Springer Verlag Berlin Heidelberg, ISSN 1432-7643 (online), ISSN 1433-7479 (print), DOI: 10.1007/s00500-013-1129-0, Vol. 18, Issue 6, pp 1093-1112

Kaiser, H. F. 1974. An index of factor simplicity. Psychometrika 39: 31–36.

Kloeden and Platen (2008). *Numerical Solutions of Stochastic Differential Equations*, pp. 11–12.

Kotsiantis, S. B.(2007). Supervised Machine Learning: A Review of Classification Techniques, Informatica, pp.249-268.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. (2006). Data preprocessing for supervised leaning, International Journal of Computer Science, Volume 1, Issue 2, pp. 111-117.

Kpigigbue N-Aabe et al, (2019). Feature reduction and prediction for wine chemical component using principal component analysis (PCA) and linear discriminant analysis (LDA). *International IJCSMC*, Vol. 8, Issue. 12, December 2019, pg.34-45

Kresimir Delac, Mislav Grgic and Sonja Grgic (2006), ‗Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set', Wiley Periodicals, Inc.

Kruskal JB. Nonmetric multidimensional scaling: A numerical method. Psychometrika. 1964; 29 (2):115–129. https://doi.org/10.1007/BF02289694

Kumar, A., Niculescu-Mizil, A., Kavukcuoglu, K., and Daume III, H. (2012). A binary classification framework for two-stage multiple kernel learning. arXiv preprint arXiv:1206.6428, Appears in Proceedings of the 29th International Conference on Machine Learning.

Larose, D. T. (2005) Dimension Reduction Methods, in Data Mining Methods and Models, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Limère, A., Laveren, E., and Van Hoof, K. (2004). A classification model for firm growth on the basis of ambitions, external potential and resources by means of decision tree

induction, Working Papers 2004027, University of Antwerp, Faculty of Applied Economics.

Lindsay I Smith, (2002). *A tutorial on Principal Components Analysis*, February 26, 2002, page 2-8.

Liu, B.S.; Chen, Y.; Shen, Y.H.; Sun, H.; Xu, X.H. A complex multi-attribute large-group decision making method based on the interval-valued intuitionistic fuzzy principal component analysis model. Soft Comput. **2014**, 18, 2149–2160. [CrossRef]

Maat, S. M., Zakaria, E., Nordin, N. M., & Meerah, T. S. M. (2011). Confirmatory factor analysis of the mathematics teachers' teaching practices instrument. *World Applied Sciences Journal*, *12*(11), 2092–2096.

Martino, L.; Luengo, D.; Míguez, J. (2012). "Efficient sampling from truncated bivariate Gaussians via Box-Muller transformation". Electronics Letters . 48 (24): 1533–1534. CiteSeerX 10.1.1.716.8683. doi:10.1049/el.2012.2816

Mehari, B., Redi-Abshiro, M., Chandravanshi, B. S., Combrinck, S., Atlabachew, M., McCrindle, R. (2016). Profiling of phenolic compounds using UPLC–MS for determining the geographical origin of green coffee beans from Ethiopia. *Journal of Food Composition and Analysis*, 45 (2016): 16–25.

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., & Saikhom, R. et al. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). *International Journal of Livestock Research, 7(5), 60-78.* http://dx.doi.org/10.5455/ijlr.20170415115235

Mohammed J. Zaki and Wagner Meira Jr. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press 2014, Chapter 7.2

Nandi, D. A. Principal component analysis in medical image processing study. *International Journal of Image Mining*, 1(1), 65-86. 2015

Nazlibilek, Sedat, Deniz Karacor, Korhan Levent Ertürk, Gokhan Sengul, Tuncay Ercan, and Fuad Aliew. "White blood cells classifications by surf image matching, pca and dendrogram." *Biomedical Research* 26, no. 4 (2015).

Nguyen LH and Holmes S (2019). Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol* 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Omucheni, D. L., Kaduki, K. A., Bulimo, W. D., & Angeyo, H. K. (2014). Application of principal component analysis to multispectral-multimodal optical image analysis for malaria diagnostics. *Malaria journal*, *13*(1), 485

Pallant, J. (2013). SPSS Survival Manual. A step by step guide to data analysis using SPSS, 4th edition. Allen & Unwin, www.allenandunwin.com/spss

Pallant, J. F. (2000). Development and validation of a scale to measure perceived control of internal states. *Journal of Personality Assessment*, *75*(2), 308–337.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Rajesh, S.; Jain, S.; Sharma, P. (2018). Inherent vulnerability assessment of rural households based on socio-economic indicators using categorical principal component analysis: A case study of Kimsar region, Uttarakhand. Ecol. Indic. **2018**, 85, 93–104. [CrossRef]

Rathod, R. R., and Momin, B. F. (2012). Performance evaluation of Outlier Detection with normalized data set, International Journal of Computer Science and Application, ISSN: 0974-0767.

Shang, L.; Wang, S. (2014). Application of improved principal component analysis in comprehensive assessment on thermal power generation units. *Power Syst. Technol*. **2014**, 38, 1928–1933. (In Chinese)

Shilton, A., and Palaniswami, M. (2008). A Unified Approach to Support Vector Machines, In B. Verma, & M. Blumenstein (Eds.), Pattern Recognition Technologies and Applications: Recent Advances, pp. 299-324.

Shirali, G.A.; Shekari, M.; Angali, K.A. (2016). Quantitative assessment of resilience safety culture using principal components analysis and numerical taxonomy: A case study in a petrochemical plant. J. Loss Prev. Process Ind. **2016**, 40, 277–284. [CrossRef]

T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. Knowledge and information systems, 10(4):453–472, 2006.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics. 5th edition.* Boston, MA: Pearson Education. Inc.

Takeda, A., Mitsugi, H., and Kanamori, T. (2012). A unified robust classification model. arXiv preprint arXiv:1206.4599.

Ting Li, et al. (2017). Adaptive scaling. arXiv:1709.00566v1 [stat.ML] 2 Sep 2017

Vasan, K.K.; Surendiran, B. Dimensionality reduction using Principal Component Analysis for network intrusion detection. Perspect. Sci. **2016**, 8, 510–512. [CrossRef]

Wold H. Estimation of Principal Components and Related Models by Iterative Least squares. In: Multivariate Analysis. New York: Academic Press; 1966. p. 391–420.

Xia, D.; Yang, S.; Li, C. Intrusion Detection System Based on Principal Component Analysis and Grey Neural Networks. In Proceedings of the 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, Wuhan, Hubei, China, 24–25 April 2010; Volume 2, pp. 142–145. [CrossRef]

Xu, J. W., Paiva, A. R., Park, I., and Principe, J. C. (2008). A reproducing kernel Hilbert space framework for information-theoretic learning, IEEE Transactions on Signal Processing, Volume 56, Issue 12, pp.5891-5902.

Y. Saad. Numerical Methods for Large Eigenvalue Problems. Halstead Press, New York, 1992.

Yee, P., and Haykin, S. (1993). Pattern classification as an ill-posed, inverse problem: a regularization approach, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93, Volume 1, pp. 597-600.

Yildiz, K.; Camurcu, Y. and Dogan, B. (2018). Comparison of Dimension Reduction Techniques on High Dimensional Datasets, *The International Arab Journal of Information Technology, Vol. 15, No. 2, March 2018*

Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining, Applied Artificial Intelligence, Volume 17, Issue 5-6,pp. 375-381.

## Appendix A: Results of one-way ANOVA and Test of Homogeneity of Variances

### a) ANOVA Test at regional and sub-regional level

| | | Region | | | | | Sub-region | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sum of Squares | df | Mean Square | F | Sig. | Sum of Squares | df | Mean Square | F | Sig. |
| 3-CQA | Between Groups | 50.130 | 3 | 16.710 | 30.825 | .000 | 77.982 | 9 | 8.665 | 32.237 | .000 |
| | Within Groups | 52.042 | 96 | .542 | | | 24.190 | 90 | .269 | | |
| | Total | 102.172 | 99 | | | | 102.172 | 99 | | | |
| 5-CQA | Between Groups | 277.032 | 3 | 92.344 | 19.793 | .000 | 303.627 | 9 | 33.736 | 7.207 | .000 |
| | Within Groups | 447.876 | 96 | 4.665 | | | 421.281 | 90 | 4.681 | | |
| | Total | 724.908 | 99 | | | | 724.908 | 99 | | | |
| 4-CQA | Between Groups | 29.267 | 3 | 9.756 | 14.581 | .000 | 61.217 | 9 | 6.802 | 18.963 | .000 |
| | Within Groups | 64.233 | 96 | .669 | | | 32.282 | 90 | .359 | | |
| | Total | 93.500 | 99 | | | | 93.500 | 99 | | | |
| 5-pCoQA | Between Groups | .017 | 3 | .006 | .255 | .857 | .451 | 9 | .050 | 2.723 | .007 |
| | Within Groups | 2.092 | 96 | .022 | | | 1.657 | 90 | .018 | | |
| | Total | 2.109 | 99 | | | | 2.109 | 99 | | | |
| 5-FQA | Between Groups | 18.941 | 3 | 6.314 | 13.040 | .000 | 33.261 | 9 | 3.696 | 10.342 | .000 |
| | Within Groups | 46.481 | 96 | .484 | | | 32.161 | 90 | .357 | | |
| | Total | 65.422 | 99 | | | | 65.422 | 99 | | | |
| 3,4-diCQA | Between Groups | 21.311 | 3 | 7.104 | 31.462 | .000 | 29.252 | 9 | 3.250 | 21.298 | .000 |
| | Within Groups | 21.676 | 96 | .226 | | | 13.735 | 90 | .153 | | |
| | Total | 42.987 | 99 | | | | 42.987 | 99 | | | |
| 3,5-diCQA | Between Groups | 85.308 | 3 | 28.436 | 12.948 | .000 | 156.661 | 9 | 17.407 | 11.232 | .000 |
| | Within Groups | 210.834 | 96 | 2.196 | | | 139.480 | 90 | 1.550 | | |
| | Total | 296.142 | 99 | | | | 296.142 | 99 | | | |
| 4,5-diCQA | Between Groups | 60.971 | 3 | 20.324 | 68.886 | .000 | 76.202 | 9 | 8.467 | 58.205 | .000 |
| | Within Groups | 28.323 | 96 | .295 | | | 13.092 | 90 | .145 | | |
| | Total | 89.294 | 99 | | | | 89.294 | 99 | | | |
| 3,5-diCQA/4,5-diCQA | Between Groups | 71.582 | 3 | 23.861 | 37.079 | .000 | 100.182 | 9 | 11.131 | 30.195 | .000 |
| | Within Groups | 61.778 | 96 | .644 | | | 33.178 | 90 | .369 | | |
| | Total | 133.360 | 99 | | | | 133.360 | 99 | | | |
| 4,5-diCQA/3,4-diCQA | Between Groups | .462 | 3 | .154 | .867 | .461 | 4.849 | 9 | .539 | 3.825 | .000 |
| | Within Groups | 17.067 | 96 | .178 | | | 12.680 | 90 | .141 | | |
| | Total | 17.529 | 99 | | | | 17.529 | 99 | | | |

**b) Test of Homogeneity of Variances for regional and sub-regional level**

| | For Region | | | | For Sub-Region | | | |
|---|---|---|---|---|---|---|---|---|
| | Levene Statistic | df1 | df2 | Sig. | Levene Statistic | df1 | df2 | Sig. |
| 3-CQA | 3.221 | 3 | 96 | .026 | 2.821 | 9 | 90 | .006 |
| 5-CQA | 4.759 | 3 | 96 | .004 | 3.176 | 9 | 90 | .002 |
| 4-CQA | 6.034 | 3 | 96 | .001 | 2.326 | 9 | 90 | .021 |
| 5-pCoQA | 2.357 | 3 | 96 | .077 | 2.308 | 9 | 90 | .022 |
| 5-FQA | 5.268 | 3 | 96 | .002 | 2.412 | 9 | 90 | .017 |
| 3,4-diCQA | 7.170 | 3 | 96 | .000 | 3.995 | 9 | 90 | .000 |
| 3,5-diCQA | 9.548 | 3 | 96 | .000 | 4.796 | 9 | 90 | .000 |
| 4,5-diCQA | 18.222 | 3 | 96 | .000 | 7.591 | 9 | 90 | .000 |
| 3,5-diCQA/4,5-diCQA | 24.400 | 3 | 96 | .000 | 8.122 | 9 | 90 | .000 |
| 4,5-diCQA/3,4-diCQA | 8.359 | 3 | 96 | .000 | 4.322 | 9 | 90 | .000 |

**Appendix B: Group Statistics at regional level**

| Region | Category | Mean | Std. Deviation | Valid N (listwise) | |
|---|---|---|---|---|---|
| | | | | Unweighted | Weighted |
| East | 3-CQA | 3.8993 | .61102 | 27 | 27.000 |
| | 5-CQA | 28.9996 | 2.03826 | 27 | 27.000 |
| | 4-CQA | 6.8000 | .50924 | 27 | 27.000 |
| | 5-pCoQA | .6989 | .15250 | 27 | 27.000 |
| | 5-FQA | 5.7000 | .71168 | 27 | 27.000 |
| | 3,4-diCQA | 2.8996 | .30615 | 27 | 27.000 |
| | 3,5-diCQA | 6.3004 | .50998 | 27 | 27.000 |
| | 4,5-diCQA | 4.9000 | .30545 | 27 | 27.000 |
| | 3,5-diCQA/4,5-diCQA | 1.2900 | .12896 | 27 | 27.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7085 | .21283 | 27 | 27.000 |
| West | 3-CQA | 2.8778 | .67295 | 18 | 18.000 |
| | 5-CQA | 33.7783 | 1.39586 | 18 | 18.000 |
| | 4-CQA | 5.8950 | .51262 | 18 | 18.000 |
| | 5-pCoQA | .6711 | .17476 | 18 | 18.000 |

| | | | | | |
|---|---|---|---|---|---|
| | 5-FQA | 4.4500 | .41122 | 18 | 18.000 |
| | 3,4-diCQA | 1.5833 | .47721 | 18 | 18.000 |
| | 3,5-diCQA | 8.6333 | 2.17817 | 18 | 18.000 |
| | 4,5-diCQA | 2.6333 | .94473 | 18 | 18.000 |
| | 3,5-diCQA/4,5-diCQA | 3.7561 | 1.63688 | 18 | 18.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7283 | .54803 | 18 | 18.000 |
| North-West | 3-CQA | 6.1833 | .41673 | 6 | 6.000 |
| | 5-CQA | 32.8333 | .75277 | 6 | 6.000 |
| | 4-CQA | 8.4000 | 1.07703 | 6 | 6.000 |
| | 5-pCoQA | .6667 | .06890 | 6 | 6.000 |
| | 5-FQA | 4.6500 | .27386 | 6 | 6.000 |
| | 3,4-diCQA | 2.6000 | .14142 | 6 | 6.000 |
| | 3,5-diCQA | 5.0167 | .46655 | 6 | 6.000 |
| | 4,5-diCQA | 3.9000 | .26077 | 6 | 6.000 |
| | 3,5-diCQA/4,5-diCQA | 1.2900 | .14422 | 6 | 6.000 |
| | 4,5-diCQA/3,4-diCQA | 1.5017 | .09559 | 6 | 6.000 |
| South | 3-CQA | 3.6694 | .83875 | 49 | 49.000 |
| | 5-CQA | 31.7959 | 2.51621 | 49 | 49.000 |
| | 4-CQA | 6.6278 | .99187 | 49 | 49.000 |
| | 5-pCoQA | .6698 | .14027 | 49 | 49.000 |
| | 5-FQA | 5.2716 | .79139 | 49 | 49.000 |
| | 3,4-diCQA | 2.0947 | .56397 | 49 | 49.000 |
| | 3,5-diCQA | 6.8302 | 1.59641 | 49 | 49.000 |
| | 4,5-diCQA | 3.5159 | .46513 | 49 | 49.000 |
| | 3,5-diCQA/4,5-diCQA | 1.9873 | .57177 | 49 | 49.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7837 | .47298 | 49 | 49.000 |
| Total | 3-CQA | 3.7398 | 1.01589 | 100 | 100.000 |
| | 5-CQA | 31.4600 | 2.70598 | 100 | 100.000 |
| | 4-CQA | 6.6487 | .97182 | 100 | 100.000 |
| | 5-pCoQA | .6777 | .14595 | 100 | 100.000 |
| | 5-FQA | 5.2021 | .81291 | 100 | 100.000 |
| | 3,4-diCQA | 2.2503 | .65895 | 100 | 100.000 |
| | 3,5-diCQA | 6.9029 | 1.72955 | 100 | 100.000 |
| | 4,5-diCQA | 3.7538 | .94972 | 100 | 100.000 |
| | 3,5-diCQA/4,5-diCQA | 2.0756 | 1.16063 | 100 | 100.000 |

| | | | | |
|---|---|---|---|---|
| 4,5-diCQA/3,4-diCQA | 1.7365 | .42079 | 100 | 100.000 |

## Appendix C: Group Statistics at sub-regional level

| Sub-Region | Category | Mean | Std. Deviation | Valid N (listwise) | |
|---|---|---|---|---|---|
| | | | | Unweighted | Weighted |
| Harar | 3-CQA | 3.8993 | .61102 | 27 | 27.000 |
| | 5-CQA | 28.9996 | 2.03826 | 27 | 27.000 |
| | 4-CQA | 6.8000 | .50924 | 27 | 27.000 |
| | 5-pCoQA | .6989 | .15250 | 27 | 27.000 |
| | 5-FQA | 5.7000 | .71168 | 27 | 27.000 |
| | 3,4-diCQA | 2.8996 | .30615 | 27 | 27.000 |
| | 3,5-diCQA | 6.3004 | .50998 | 27 | 27.000 |
| | 4,5-diCQA | 4.9000 | .30545 | 27 | 27.000 |
| | 3,5-diCQA/4,5-diCQA | 1.2900 | .12896 | 27 | 27.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7085 | .21283 | 27 | 27.000 |
| Jimma A | 3-CQA | 2.5000 | .10000 | 3 | 3.000 |
| | 5-CQA | 35.3333 | .57735 | 3 | 3.000 |
| | 4-CQA | 5.6333 | .15275 | 3 | 3.000 |
| | 5-pCoQA | .9200 | .04000 | 3 | 3.000 |
| | 5-FQA | 4.5000 | .30000 | 3 | 3.000 |
| | 3,4-diCQA | 1.2333 | .05774 | 3 | 3.000 |
| | 3,5-diCQA | 8.8333 | .70238 | 3 | 3.000 |
| | 4,5-diCQA | 1.9667 | .15275 | 3 | 3.000 |
| | 3,5-diCQA/4,5-diCQA | 4.4900 | .03606 | 3 | 3.000 |
| | 4,5-diCQA/3,4-diCQA | 1.5967 | .08737 | 3 | 3.000 |
| Jimma B | 3-CQA | 4.2000 | .40000 | 3 | 3.000 |
| | 5-CQA | 32.0000 | 1.00000 | 3 | 3.000 |
| | 4-CQA | 6.9000 | .40000 | 3 | 3.000 |
| | 5-pCoQA | .5133 | .02082 | 3 | 3.000 |
| | 5-FQA | 4.1667 | .35119 | 3 | 3.000 |
| | 3,4-diCQA | 2.0667 | .25166 | 3 | 3.000 |

| | | | | | |
|---|---|---|---|---|---|
| | 3,5-diCQA | 5.6333 | .75056 | 3 | 3.000 |
| | 4,5-diCQA | 4.1000 | .30000 | 3 | 3.000 |
| | 3,5-diCQA/4,5-diCQA | 1.3700 | .08000 | 3 | 3.000 |
| | 4,5-diCQA/3,4-diCQA | 1.9900 | .10583 | 3 | 3.000 |
| Kaffa | 3-CQA | 2.6000 | .31756 | 10 | 10.000 |
| | 5-CQA | 34.0010 | 1.05451 | 10 | 10.000 |
| | 4-CQA | 5.7010 | .21126 | 10 | 10.000 |
| | 5-pCoQA | .6090 | .12662 | 10 | 10.000 |
| | 5-FQA | 4.6000 | .42293 | 10 | 10.000 |
| | 3,4-diCQA | 1.6000 | .52772 | 10 | 10.000 |
| | 3,5-diCQA | 9.8000 | 1.89898 | 10 | 10.000 |
| | 4,5-diCQA | 2.5000 | .84279 | 10 | 10.000 |
| | 3,5-diCQA/4,5-diCQA | 4.3350 | 1.60264 | 10 | 10.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7120 | .72808 | 10 | 10.000 |
| Wolega | 3-CQA | 2.8500 | .07071 | 2 | 2.000 |
| | 5-CQA | 33.0000 | 1.41421 | 2 | 2.000 |
| | 4-CQA | 5.7500 | .21213 | 2 | 2.000 |
| | 5-pCoQA | .8450 | .10607 | 2 | 2.000 |
| | 5-FQA | 4.0500 | .21213 | 2 | 2.000 |
| | 3,4-diCQA | 1.3000 | .00000 | 2 | 2.000 |
| | 3,5-diCQA | 7.0000 | .14142 | 2 | 2.000 |
| | 4,5-diCQA | 2.1000 | .14142 | 2 | 2.000 |
| | 3,5-diCQA/4,5-diCQA | 3.3400 | .15556 | 2 | 2.000 |
| | 4,5-diCQA/3,4-diCQA | 1.6150 | .10607 | 2 | 2.000 |
| Benishangul | 3-CQA | 6.5000 | .20000 | 3 | 3.000 |
| | 5-CQA | 33.0000 | 1.00000 | 3 | 3.000 |
| | 4-CQA | 9.3000 | .40000 | 3 | 3.000 |
| | 5-pCoQA | .6400 | .09000 | 3 | 3.000 |
| | 5-FQA | 4.5667 | .20817 | 3 | 3.000 |
| | 3,4-diCQA | 2.6333 | .20817 | 3 | 3.000 |
| | 3,5-diCQA | 4.8000 | .20000 | 3 | 3.000 |
| | 4,5-diCQA | 4.1000 | .10000 | 3 | 3.000 |

| | | | | | |
|---|---|---|---|---|---|
| | 3,5-diCQA/4,5-diCQA | 1.1700 | .02000 | 3 | 3.000 |
| | 4,5-diCQA/3,4-diCQA | 1.5633 | .09292 | 3 | 3.000 |
| Finote-Selam | 3-CQA | 5.8667 | .30551 | 3 | 3.000 |
| | 5-CQA | 32.6667 | .57735 | 3 | 3.000 |
| | 4-CQA | 7.5000 | .55678 | 3 | 3.000 |
| | 5-pCoQA | .6933 | .04041 | 3 | 3.000 |
| | 5-FQA | 4.7333 | .35119 | 3 | 3.000 |
| | 3,4-diCQA | 2.5667 | .05774 | 3 | 3.000 |
| | 3,5-diCQA | 5.2333 | .60277 | 3 | 3.000 |
| | 4,5-diCQA | 3.7000 | .20000 | 3 | 3.000 |
| | 3,5-diCQA/4,5-diCQA | 1.4100 | .09165 | 3 | 3.000 |
| | 4,5-diCQA/3,4-diCQA | 1.4400 | .05292 | 3 | 3.000 |
| Sidama SA | 3-CQA | 2.7000 | .42208 | 10 | 10.000 |
| | 5-CQA | 30.9990 | 3.16157 | 10 | 10.000 |
| | 4-CQA | 5.6990 | .63296 | 10 | 10.000 |
| | 5-pCoQA | .7290 | .17792 | 10 | 10.000 |
| | 5-FQA | 4.5010 | .52817 | 10 | 10.000 |
| | 3,4-diCQA | 1.7010 | .31586 | 10 | 10.000 |
| | 3,5-diCQA | 7.1010 | .94695 | 10 | 10.000 |
| | 4,5-diCQA | 2.7990 | .31719 | 10 | 10.000 |
| | 3,5-diCQA/4,5-diCQA | 2.5770 | .51053 | 10 | 10.000 |
| | 4,5-diCQA/3,4-diCQA | 1.6890 | .30686 | 10 | 10.000 |
| Sidama SB | 3-CQA | 4.2000 | .61123 | 29 | 29.000 |
| | 5-CQA | 32.0000 | 2.03577 | 29 | 29.000 |
| | 4-CQA | 7.1997 | .81450 | 29 | 29.000 |
| | 5-pCoQA | .6700 | .13213 | 29 | 29.000 |
| | 5-FQA | 5.7000 | .61182 | 29 | 29.000 |
| | 3,4-diCQA | 2.4007 | .50945 | 29 | 29.000 |
| | 3,5-diCQA | 6.2993 | 1.42524 | 29 | 29.000 |
| | 4,5-diCQA | 3.6997 | .30576 | 29 | 29.000 |
| | 3,5-diCQA/4,5-diCQA | 1.7117 | .40959 | 29 | 29.000 |
| | 4,5-diCQA/3,4-diCQA | 1.6207 | .41185 | 29 | 29.000 |

| | | | | | |
|---|---|---|---|---|---|
| Yirgachefe | 3-CQA | 3.1000 | .31612 | 10 | 10.000 |
| | 5-CQA | 32.0010 | 3.16089 | 10 | 10.000 |
| | 4-CQA | 5.8980 | .42234 | 10 | 10.000 |
| | 5-pCoQA | .6100 | .10541 | 10 | 10.000 |
| | 5-FQA | 4.8000 | .63122 | 10 | 10.000 |
| | 3,4-diCQA | 1.6010 | .21074 | 10 | 10.000 |
| | 3,5-diCQA | 8.0990 | 1.89666 | 10 | 10.000 |
| | 4,5-diCQA | 3.7000 | .21182 | 10 | 10.000 |
| | 3,5-diCQA/4,5-diCQA | 2.1970 | .53012 | 10 | 10.000 |
| | 4,5-diCQA/3,4-diCQA | 2.3510 | .34359 | 10 | 10.000 |
| Total | 3-CQA | 3.7398 | 1.01589 | 100 | 100.000 |
| | 5-CQA | 31.4600 | 2.70598 | 100 | 100.000 |
| | 4-CQA | 6.6487 | .97182 | 100 | 100.000 |
| | 5-pCoQA | .6777 | .14595 | 100 | 100.000 |
| | 5-FQA | 5.2021 | .81291 | 100 | 100.000 |
| | 3,4-diCQA | 2.2503 | .65895 | 100 | 100.000 |
| | 3,5-diCQA | 6.9029 | 1.72955 | 100 | 100.000 |
| | 4,5-diCQA | 3.7538 | .94972 | 100 | 100.000 |
| | 3,5-diCQA/4,5-diCQA | 2.0756 | 1.16063 | 100 | 100.000 |
| | 4,5-diCQA/3,4-diCQA | 1.7365 | .42079 | 100 | 100.000 |

## Appendix D: MATLAB codes

```
% PCA Matlab code

function PCA8(data)
% Calculating the lower dimensional space of the PCA using covariance matrix.
% Sample dataset read from Excel file
   data=xlsread('coffeedata.xlsx','sheet20','D1:L101')'; % (8x100)-dimensional dataset
   [r,c] = size(data);
% Compute the mean of all data points for each dimension
   Xmean= mean(data,2)*ones(1,c); % matrix whose rows are filled with the mean of data for that
row
   Xstd=std(data,0,2)*ones(1,c); % matrix whose rows are filled with the standard deviation of data
for that row
   Xmax= max(data')'*ones(1,c); % matrix whose rows are filled with the maximum of data for that
row
   Xmin= min(data')'*ones(1,c); % matrix whose rows are filled with the minimum of data for that
row
   % Normalizing data so that each category is a different row and
   % Feature Scaling is a necessary step for data pre-processing
   method= input('Enter Scaling Method: 1-for Standardization(Z-score), 2-for Pareto-scaling, Any
other # for New Scale: ')
   if method==1
      % 1)Standardization (Z-score transform) on centered data to have a unit variance
      data=(data-Xmean)./Xstd; % Standardization (Z-score transform) on centered data to make
features'a variance to be 1
   elseif method==2
      % 2)Pareto-Scaling on centered data to reduce the impact of variance
      data=(data-Xmean)./sqrt(Xstd); % Pareto-Scaling on centered data to reduce the impact of
variance
   else
      data=(data-Xmean)./(Xmax-Xmin);
   end
% Compute the covariance matrix using the normalized data
   C= (data*data')/(c-1);
% Do an eigendecomposition
   % EIG Eigenvalues and eigenvectors.
   % [V,D] = EIG(X) produces a diagonal matrix D of eigenvalues and a full
   % matrix V whose columns are the corresponding eigenvectors so that X*V = V*D.
   [PC,V]= eig(C);  %Compute the eigenvectors and eigenvalues of the Covariance matrix, C so that
C*V = V*D, or C=V*D*V'
% Extract diagonal of Variance matrix as vector
   V= diag(V); % Variances along eigenvectors
% Sort the variances in decreasing order
   [~,index]= sort(V,'descend');
   V= V(index); % variances (eigenvales) in descending order
   Explained=100*V/sum(V); %variances of all individual principal components
   PC= PC(:,index); % the Coefficient matrix of the Principal Components
% Scree plot of variances
   figure (1)
   bar(V(1:r))
   xlabel('Components')
   ylabel('Variance')
% Use the following to display the PCA (Outputs) using tables and graphs
```

```
% factors- Factors (Principal Components)
% vbls- Variables (features) of the dataset
% obsrbs- Observations (individuals) of the dataset
   factors={'PC 1','PC 2','PC 3','PC 4','PC 5','PC 6','PC 7','PC 8'};
   vbls={'3-CQA','5-CQA','4-CQA','5-pCoQA','5-FQA','3,4-diCQA','3,5-diCQA','4,5-diCQA'};
   %%
disp('====================================================')
fprintf('Variance:'), disp(V');
fprintf('Explained:'), disp(Explained');
%Cumulative sum of variances
fprintf('Cumu_Sum: ')
   cumsm=(cumsum(V)./sum(V))*100; disp(cumsm');
disp('----------------------------------------------------')
% The Principal Components (PCs)(or Column Vectors:)
fprintf('Principal Components (Columns):\n')
disp('----------------------------------------------------')
   pcts=[{'Variable'} factors(1:8);vbls' num2cell(PC(:,1:8))];
   disp(pcts)
   %xlswrite('PCA_files',pcts,'sheet1','A5')
disp('----------------------------------------------------')
% Factor Loadings:
disp('Factor Loadings: ')
disp('----------------------------------------------------')
for i=1:r
   FL(:,i)=sqrt(V(i))*PC(:,i); % compute Factor Loadings for each variable
end
   fLoads=[{'Variable'} factors(1:8);vbls' num2cell(FL(:,1:8))];
   disp(fLoads)
   %xlswrite('PCA_files',fLoads,'sheet1','A15')

  % Rescaling component loadings
   k=3;
      for i=1:r
          SqL(i)=FL(i,1)^2+FL(i,2)^2+FL(i,3)^2;
      end
      for i=1:r
         for j=1:k
             FL(i,j)=(1/sqrt(SqL(i)'))*FL(i,j);
         end
      end
     fLoads=[{'Variable'} factors(1:k);vbls' num2cell(FL(:,1:k))];
           % Normalizing Vectors
              for j=1:k

VecL(j)=sqrt(FL(1,j)^2+FL(2,j)^2+FL(3,j)^2+FL(4,j)^2+FL(5,j)^2+FL(6,j)^2+FL(7,j)^2+FL(8,j)^2);
              end
              %disp(VecL)
              for i=1:r
                 for j=1:k
                    FL(i,j)=(1/VecL(j))*FL(i,j);
                 end
              end
              disp(FL)
              fLoads=[{'Variable'} factors(1:k);vbls' num2cell(FL(:,1:k))];
              disp(fLoads(:,1:k+1))
```

```
    % Factor rotation
        [L,T]=rotatefactors(FL,'method','Varimax');
        FL=L*L';
        disp(FL(:,1:k))

% Squared Loadings (Cosines) of the Variables:
disp('Squared Loadings: ')
for i=1:k
   SL(:,i)=FL(:,i).^2; % Computes Squared loadings (Cosines) for each variables
end
   sqLoads=[{'Variable'} factors(1:k);vbls' num2cell(SL(:,1:k))];
   %disp(sqLoads)
%     xlswrite('pcalda_results',sqLoads,'sheet1','L44')
% Contribution of the Variables (%):
disp('Contribution of the Variables: ')
for i=1:k
  CV(:,i)=100*SL(:,i)./V(i);% Computes contribution of variables for each factor
end
   contrVar=[{'Variable'} factors(1:k);vbls' num2cell(CV(:,1:k))];
   disp(contrVar(:,1:k+1))
%     xlswrite('pcalda_results',contrVar,'sheet1','L57')
% Factor (PC) scores for each observation
        score= data'*FL;
       %  disp(score)
disp('Factor Scores: ')
   factor_sc=[factors(1:k); num2cell(score(:,1:k))];
   %disp(factor_sc)
   %xlswrite('pcalda_results',factor_sc,'sheet5','A2')
% The Squared Cosines of the individuals (observation)
disp('Squared Cosines: ')
   score2=score.^2;
   for i=1:c
      Sscore(i,:)=score2(i,:)./sum(score2(i,:)); % Computes Squared Cosines(Scores) for each
observation
   end
   squared_cos=[factors(1:k);num2cell(Sscore(:,1:k))];
   %disp(squared_cos)
   %xlswrite('PCA_files',squared_cos,'sheet1','A100')
disp('% Contribution of individual observation in each component
disp('Contribution of the Observation (%): ')
   for j=1:k
      COb(:,j)=100*score2(:,j)./sum(score2(:,j));
   end
   Contrib_ob=[factors(1:k);num2cell(COb(:,1:k))];
   %disp(Contrib_ob)
   %xlswrite('PCA_files',Contrib_ob,'sheet1','A155')
FLs=FL; PCs=score;
 figure (3) % 3D biplot of Loadings and Scores for (PC-1, PC-2, PC-3)
% FLs- Factor Loadings
% vbls- variables (features) defined above
biplot(FLs(:,1:3),'Scores',PCs(:,1:3),'Color','r','Marker','*','VarLabels',vbls);
      title('The biplot of Loadings & scores for the 1st two PCs')
   xlabel('PC 1')
   ylabel('PC 3')
figure (4)  % 2-dimension plot of Variables on the first two PCs(PC1 and PC2)
```

```
% FLs- Factor Loadings defined above
plot(FLs(1,1),FLs(1,2),'b.',FLs(2,1),FLs(2,2),'b.',FLs(3,1),FLs(3,2),'b.',FLs(4,1),FLs(4,2),'b.',FLs(5,1),
FLs(5,2),'b.',FLs(6,1),FLs(6,2),'b.',FLs(7,1),FLs(7,2),'b.',FLs(8,1),FLs(8,2),'b.','markersize',20)
   title('2D Plot of Variables on the first two PCs')
   xlabel('Principal Component 1'), ylabel('Principal Component 2')
   grid on
   for i=1:r
      text(FLs(i,1)+0.01,FLs(i,2)-0.01,vbls(i));
   end
figure (6) % the biplot of Loadings and Scores for the first two PCs (PC-1 and PC-2)
% FLs- Factor Loadings
% vbls- variables (features) defined above
biplot(FLs(:,1:2),'Scores',PCs(:,1:2),'Color','r','Marker','*','VarLabels',vbls);
   title('The biplot of Loadings & scores for the 1st two PCs')
   xlabel('PC 1')
   ylabel('PC 2'

score=PCs';
figure (8)
plot(score(1,1:27),score(2,1:27),'*','MarkerEdgeColor','g','LineWidth',2,'MarkerSize',12); hold on
plot(score(1,28:45),score(2,28:45),'v','MarkerEdgeColor','r','LineWidth',2,'MarkerSize',10); hold on
plot(score(1,46:51),score(2,46:51),'d','MarkerEdgeColor','b','LineWidth',2,'MarkerSize',10); hold on
plot(score(1,52:100),score(2,52:100),'k+','MarkerEdgeColor','k','LineWidth',2,'MarkerSize',12)
   %grid on
   title('The Projected Data of the 4 Regions on 1st two PCs')
   xlabel('PC 1')
   ylabel('PC 2')
   legend('East','West','Northwest','South')
% Robustness of PCA space
   robust2d=(sum(V(1:2))/sum(V))*100;
   robust3d=(sum(V(1:3))/sum(V))*100;
   disp('Robustness of PCA space')
   disp('2D (1st two PCs) (%)'), disp(robust2d);
   disp('3D (1st three PCs) (%)'), disp(robust3d);
end

%%%%%%%
```

**% LDA Matlab code**

```
function LDA8(X,K)
%Matlab code for LDA on green_coffee dataset
% Sample dataset read from Excel file
   data=xlsread('coffeedata.xlsx','sheet20','D1:N101')'; % (10x100)-dimensional dataset
   [d,n] = size(data);
X=data';
q=3;
[d,n]= size(X');   % d features, n patterns of the data samples
L(1)=27; L(2)=45; L(3)=51; L(4)=100;
n1=27; n2=18; n3=6; n4=49;
% (a) - Covariance Matrices
X1=X(1:L(1),:); X2=X(1+L(1):L(2),:); X3=X(1+L(2):L(3),:); X4=X(1+L(3):L(4),:);
m1=mean(X1); m2=mean(X2); m3=mean(X3); m4=mean(X4);
mu=mean(X); T=cov(X);
W1=cov(X1); W2=cov(X2); W3=cov(X3); W4=cov(X4);
Sw=(n1-1)*W1+(n2-1)*W2+(n3-1)*W3+(n4-1)*W4;
```

```
Sb=(n1-1)*(m1-mu)'*(m1-mu)+(n2-1)*(m2-mu)'*(m2-mu)+(n3-1)*(m3-mu)'*(m3-mu)+(n4-1)*(m4-
mu)'*(m4-mu);
St=Sw+Sb;
% Computing Eigenvalues & eigenvectors of (Swl)^-1*Sbl
s=0.0001;
A=(inv(Swl+s*eye(d)))*(Sbl+eps);    % gives better separability than A= Swl'*Sbl;
 [U D]=eig(A);
Lambda=diag(D);
[~, index]=sort(Lambda,'descend');
eigval= Lambda(index);
   eigval= real(eigval);
   Explained= 100*eigval/sum(eigval);
   cumusum= (cumsum(eigval)./sum(eigval))*100;
   disp(eigval')
   disp(Explained')
   disp(cumusum')
% Computing the LDA subspace
K=q;
Wlda=zeros(K,d); % initiate LDA subspace
for i=1:K
   Wlda(i,:)=real(U(:,index(i))');
end
% The Projection matrix from PCA+LDA is
Pl= Wpca*Wlda';
Y=(X*Pl); % The first K discriminative components
X1=Y(1:L(1),1); Y1=Y(1:L(1),2); Z1=Y(1:L(1),3);
X2=Y(1+L(1):L(2),1); Y2=Y(1+L(1):L(2),2); Z2=Y(1+L(1):L(2),3);
X3=Y(1+L(2):L(3),1); Y3=Y(1+L(2):L(3),2); Z3=Y(1+L(2):L(3),3);
X4=Y(1+L(3):L(4),1); Y4=Y(1+L(3):L(4),2); Z4=Y(1+L(3):L(4),3);
figure (1)
plot(X1,Y1,'d',X2,Y2,'O',X3,Y3,'*',X4,Y4,'^'); grid
title('PCA+LDA: Dataset Projection on the first two LDs')
xlabel('Function 1'),ylabel('Function 2')
legend('East','West','Northwest','South')
% Next codes compute  LDA
s=0.0001;
C=(inv(Sw+s*eye(d)))*(Sb+eps);
% (b) - Compute Eigenvalues of W^{-1}B
[U D]=eig(C);
Lambda=diag(D);
[~, index]=sort(Lambda,'descend');
eigval= Lambda(index);
   eigval= real(eigval);
   Explained= 100*eigval/sum(eigval);
   cumusum= (cumsum(eigval)./sum(eigval))*100;
% 2d and 3d LDA Projection for green_coffee data set
     K=q;
     Xproj=zeros(K,d); % initiate a projection matrix
     for i=1:K
        Xproj(i,:)=real(U(:,index(i))');
     end
     disp(Xproj')
     Y=(Xproj*X')'; % DA projection on function 1 and 2
      xlswrite('pcalda_results',Y,'sheet11','A2')
```

```
    X1=Y(1:L(1),1); Y1=Y(1:L(1),2); Z1=Y(1:L(1),3);
    X2=Y(1+L(1):L(2),1); Y2=Y(1+L(1):L(2),2); Z2=Y(1+L(1):L(2),3);
    X3=Y(1+L(2):L(3),1); Y3=Y(1+L(2):L(3),2); Z3=Y(1+L(2):L(3),3);
    X4=Y(1+L(3):L(4),1); Y4=Y(1+L(3):L(4),2); Z4=Y(1+L(3):L(4),3);
End
```